

Sheffield Hallam University

Artificial Intelligence Facial Expression Recognition for Early Prediction of Human Health Deterioration

AL-TEKREETI, Zeena Sabah Ismaeel

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/34872/>

A Sheffield Hallam University thesis

This thesis is protected by copyright which belongs to the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Please visit <https://shura.shu.ac.uk/34872/> and <http://shura.shu.ac.uk/information.html> for further details about copyright and re-use permissions.

Artificial Intelligence Facial Expression Recognition for Early Prediction of Human Health Deterioration

Zeena Sabah Ismaeel Al-Tekreeti

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF
SHEFFIELD HALLAM UNIVERSITY
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

May 2024

Declaration

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.
2. None of the material contained in the thesis has been used in any other submission for an academic award.
3. I am aware of and understand the University's policy on plagiarism and certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.
4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.
5. The word count of the thesis is 50,000.

Name	Zeena Sabah Ismaeel Al-Tekreeti
Award	PhD
Date of Submission	May 2024
Faculty	The Department of Engineering and Mathematics Industry and Innovation Research Institute.
Director(s) of Studies	Professor Marcos Rodrigues

Dedication

This research work is dedicated to:

- ❖ My parents, who have raised me to be the person I am today. None of this would have been possible without your unending love, support and encouragement, which have served as a foundation for me during the best and worst times.
- ❖ To my mother, who has left us with her body, but whose soul still flutters in the sky of my life.
- ❖ To my father, whoever has left our world, and whose advice still guides me.
- ❖ My lovely kids Yousif and Elaine, your kind support, sacrifices and good wishes were more worthy than you could imagine.
- ❖ To all those from whom I have received advice and support, thank you for your love and the ongoing encouragement over the years.

I dedicate to you the summary of my scientific effort.

Zeena

Acknowledgements

I would like to thank some people, where this work would never have been possible without their support and guidance.

I wish to express my sincere thanks and deep gratitude to my Director of Study Professor Marcos Rodrigues for his invaluable guidance, constant encouragement, inspiration, and assistance throughout this PhD project. You have always encouraged me to follow my own path and to constantly strive to become a better researcher. I am fortunate to have you as my mentor.

I would also like to express my appreciation to Jeronimo Moreno-Cuesta and Maria Isabel Madrigal, who works at the North Middlesex University Hospital, for giving us the problem and their assistance throughout this research project.

Special thanks should also be given to MERI staff (Dr. Da Costa Abreu Marjory, Jennifer Sturch, Alexandra Preston, Amy) and my PhD colleagues at MERI for their kind support whenever required.

I would also like to express my special thanks to The Ministry of Higher Education and Scientific Research in Iraq, the University of Technology, and the Iraqi Cultural Attache for their assistance and support throughout this research project.

Finally, I am indebted to my sponsor Sheffield Hallam University for financial support and for giving me this opportunity to complete my PhD.

Author

Abstract

Facial expressions are a universally recognised means of conveying internal emotional states across diverse human cultural and ethnic groups. Recent advances in understanding people's emotions expressed through verbal and non-verbal communication are particularly noteworthy in the clinical context for the assessment of patients' health and well-being. Facial expression recognition (FER) plays an important and vital role in healthcare, providing communication with a patient's feelings and allowing the assessment and monitoring of mental and physical health conditions. However, the subtle and rapid nature of facial expressions poses a challenge to swift recognition and interpretation. Previous research collaboration between North Middlesex Hospital and the GMPR group has demonstrated for the first time that human recognised patterns of facial action units can be used to predict admission to intensive care. This thesis shows that automatic machine learning methods may predict health deterioration accurately and robustly, independent of human subjective assessment. Methods are developed to create a facial database mimicking the underlying muscular structure of the face, whose Action Unit motions can then be transferred to human face images, thus displaying animated expressions of interest. To detect and recognize expressions, five models are proposed and tested. The first model combined face detection method with a 1D Convolutional Neural Network (1D-CNN), using raw generated data coordinates as input. Results show 99.74% accuracy in predicting patient deterioration. The second model combines 1D-CNN with Long Short-Term Memory (LSTM) with different data pre-processing methods with an overall accuracy of 99.89%. The third and fourth models, based on Random Forest and Support Vector Machine methods yield accuracies of 100% and 60% respectively. Finally, the Transformer model yields a low accuracy of 20%. The main contributions to knowledge from this thesis can be summarized as 1) the generation of visual datasets mimicking real-life samples of facial expressions indicating health deterioration; 2) to improve understanding and communication with patients at risk of deterioration through facial expression analysis, and 3) developing state-of-the-art models to recognize such facial expressions based on simulated facial expressions. Hence, the significance of the investigation and prediction model designs is to directly support clinical systems in detecting and assessing early signs of health deterioration directly from the analysis of patients' facial expressions. As such, the outcomes of this PhD thesis may help to improve assessment of health deterioration by introducing real-time, health trend analysis and early warning systems to support timely interventions.

Table of Content

Contents

Dedication	iii
Acknowledgements.....	iv
Abstract	v
Table of Content	vi
List of Table	ix
List of Figures	x
List of Abbreviations	xiv
Chapter 1 Introduction.....	1
1.1 Background	1
1.2 Related Work.....	5
1.3 The Significance of this Research.....	7
1.4 Motivation.....	8
1.5 Problem Statement.....	9
1.6 Aim and Objectives	9
1.7 Expected Contribution to Knowledge	11
1.8 Publication.....	12
1.9 Methodology.....	11
1.10 Thesis Structure	18
Chapter 2 Facial Expression: A Literature Review	19
2.1 Introduction.....	19
2.2 Facial Expressions (FEs).....	22
2.2.1 Micro-Expressions	23
2.2.2 Macro expressions.....	25
2.2.3 The Design of an Automatic Micro-FER.....	27
2.3 Facial Expression Recognition (FER).....	29
2.4 Automatic Micro Facial Expression Recognition (AFER)	31
2.4.1 Pre-Processing	35
2.4.2 Feature Extraction	37
2.5 Upper and Lower Parts of the Face	42
2.6 Face Detection using MediaPipe.....	44
2.7 Facial Landmarks	46
2.8 Conclusion	50

Chapter 3 Introduction to Machine Learning	51
3.1 Introduction to Machine Learning and Data analysis.....	51
3.2 Machine Learning Models	58
3.2.1 Support Vector Machine (SVM)	59
3.2.2 Random Forest	62
3.3 Deep Learning Models.....	64
3.3.1 Improving Facial Expression Recognition via Deep Learning	65
3.3.2 Techniques to Improve Performance of DL Models.....	66
3.3.3 The Convolution Neural Network (CNN).....	67
3.3.4 Recurrent Neural Network and Long Short-Term Memory.....	70
3.3.5 Convolution LSTM	72
3.3.6 Transformers	74
3.4 Some Important DL Model Activation Functions	78
3.5 Genetic Algorithm (GA) for Hyperparameters Selection	79
3.6 Conclusions.....	80
Chapter 4 Datasets Description and Analysis	82
4.1 Introduction.....	82
4.2 AUs Analysis	86
4.3 Generating the Dataset.....	91
4.3.1 Generating a Facial Expression Dataset via Avatars	91
4.3.2 Generating Data-Driven Facial Expressions using FACS	94
4.4 The Transfer of Facial Expressions to Static Real Faces using the First Order Motion Model (FOMM).....	104
4.5 Conclusion	108
Chapter 5 Facial Expression Classification Using Random Forest and Support Vector Machines	110
5.1 The Proposed Method.....	110
5.2 Data Pre-Processing and Feature Extraction	113
5.2.1 Pre-Processing Techniques	113
5.2.2 Feature Extraction	115
5.3 The Random Forest Classifier	118
5.3.1 RF Model Results and Evaluation.....	118
5.3.2 Real-Time Model Evaluation.....	123
5.4 Support Vector Machine (SVM).....	126
5.5 Conclusion	130

Chapter 6 Facial Expression Classification Using Deep Learning	130
6.1 The Proposed Method using Deep learning Models.....	130
6.2 A 1D-CNN Model to Predict Patients at Risk of Deterioration.....	134
6.2.1 Data Pre-processing and Design of a 1D-CNN Model	136
6.2.2 Results and Evaluation of the 1D-CNN Model.....	141
6.3 A 1D-ConvLSTM Model to Predict Patients at Risk of Deterioration	152
6.3.1 Data Pre-Processing and Design of a 1D-ConvLSTM Model	153
6.3.2 Results and Evaluation of the 1D-ConvLSTM Model.....	156
6.4 Transformers.....	166
6.5 Model Comparative Analysis.....	170
6.6 Conclusions.....	175
Chapter 7 Conclusions and Future Directions	176
7.1 Conclusions.....	176
7.2 Findings and the Impact of the Project.....	178
7.3 Recommendation for Future Work.....	179
References	180
APPENDIX A. PUBLISHED PAPER.....	203

List of Table

Table 4.1 Action Units of five expressions at risk of deterioration and their relevant facial muscles.....	92
Table 4.2 Combination of AUs of each class to form facial expressions of participants at risk of deterioration & Number of generated videos	103
Table 5.1 Evaluation Metrics for each Class & The Total Mean Performance of RF Model.	120
Table 5.2 Evaluation Metrics for each Class & The Total Mean Performance of SVM Model.	128
Table 6.1 Evaluation Metrics for each Class % The Total Mean Performance of 1D-CNN Model.	148
Table 6.2 Evaluation Metrics for each Class & The Total Mean Performance of 1D-ConvLSTM Model.....	161
Table 6.3 Accuracy of the models using generated database PRD-FE.....	171
Table 6.4 The overall error rate of the designed model of predicting FEs of patients at risk of Deterioration	171

List of Figures

Figure 1.1 The waterfall model with feedback.	13
Figure 1.2 Methodology of Proposed AFER system.	16
Figure 1.3 Virtual Simulation Environment depicts Intelligent ICU includes the proposed AI-VIEW system.	17
Figure 2.1 Main facial expression muscles of face and key facial areas division	20
Figure 2.2 Main facial expression muscles of Areas and facial muscles that reflect deterioration critical condition of patients are spotted with red circles.	21
Figure 2.3 Intensity motion difference between micro and macro expression	27
Figure 2.4 Facial expression areas that reveal if the patient is under deterioration or not.	28
Figure 2.5 Basic framework of automatic facial expression recognition.	35
Figure 2.6 Some of main regions of interest of facial landmarks	47
Figure 2.7 Two different facial landmarks models.	49
Figure 3.1 Types of Machine Learning Algorithms.	53
Figure 3.2 Euler Diagram showing the AI hierarchy.	54
Figure 3.3 K-Fold Cross Validation.....	58
Figure 3.4 An Example showing a Linear SVM Classifier	60
Figure 3.5 An Example showing a non-Linear SVM Classifier	61
Figure 3.6 Random Forest Model.	64
Figure 3.7 Drop out Technique.	66
Figure 3.8 General Architecture 1D-CNN based on.....	69
Figure 3.9 Basic architecture of RNN based on.....	70
Figure 3.10 Inner structure of LSTM cell	72
Figure 3.11 The structure of ConvLSTM.	73
Figure 3.12 Vision Transformer Architecture.....	78
Figure 4.1 Methodology of generating dataset (PRD-FE).	85
Figure 4.2 Illustrates the patient go through the deterioration stage.	88
Figure 4.3 Facial expression areas that reveal if the patient is under deterioration or not.	89

Figure 4.4 Expressive images and their active AU coding based on FACS.	90
Figure 4.5 Sample of coded AU with FACS.	94
Figure 4.6 FACS Human user interface.	96
Figure 4.7 Progress of facial motion intensity.	96
Figure 4.8 Characteristics of one Action Unit defined on the timeline.	97
Figure 4.9 Sequence of images produced using the FANT plugin as an example of transition from anger expression to surprise expression.	98
Figure 4.10 An analysis of the chronological sequence and layout of AUs intensity	100
Figure 4.11 Description of progress of the synchronized AUs based on timeline.	102
Figure 4.12 Scene editor and lighting possibilities.	102
Figure 4.13 Five classes along with the combination of Action Units.	105
Figure 4.14 Frames of Video Sample After Utilizing FOMM to transfer facial expressions from avatars.	107
Figure 4.15 Bar chart depicts number of generated videos in each class of 5 classes. ...	108
Figure 4.16 Samples of five classes.	109
Figure 5.1 The proposed system flowchart based on facial landmarks as feature extraction method along with RF or SVM as classifiers.	111
Figure 5.2 Main Stages of AFER based on feature extraction and Machine Learning. ...	113
Figure 5.3 Facial frames samples for each class after pre-processed using a face mesh as face detection technique.	114
Figure 5.4 Four designed Triangles to extract features of head tilt direction.	115
Figure 5.5 Overview of the proposed method for facial expression recognition.	118
Figure 5.6 The Confusion Matrix of RF Classifier.	118
Figure 5.7 The Classification Report of RF Model.	119
Figure 5.8 The Evaluation Metrics of RF Model.	119
Figure 5.9 The RF classifier accurately classified unseen data into 5 categories.	122
Figure 5.10 Shows real-time model performance on author face.	124
Figure 5.11 Other FEs examined by the model.	125
Figure 5.12 The Confusion Matrix of SVM Model.	126

Figure 5.13 The Evaluation Metrics of SVM Model.....	127
Figure 5.14 The Classification Report of SVM Model.....	127
Figure 6.1 Number and ratio of Samples in each class for the whole dataset.	131
Figure 6.2 Number of Samples in training and test dataset.	132
Figure 6.3 Number of Training Samples Before and After Oversampling Method.	133
Figure 6.4 Phases of AFER based on DNNs	134
Figure 6.5 CNN-based FER method.....	135
Figure 6.6 Preprocessing Dataset by Face Detection & then convert images to grey scale.	137
Figure 6.7 The design of the 1-D CNN model.....	138
Figure 6.8 Layers of proposed 1D-CNN model.....	139
Figure 6.9 The stages of 1D-CNN based FER.....	140
Figure 6.10 The Evaluation Metrics of 1D-CNN Model Performance.....	141
Figure 6.11 Loss, Mean Square Error, and Mean Absolute Error of 1D-CNN Model....	142
Figure 6.12 The Confusion Matrix of 1D-CNN Model.....	143
Figure 6.13 The Evaluation Metrics of 1D-CNN Model.....	144
Figure 6.14 The Classification Report of 1D-CNN Model.....	145
Figure 6.15 Evaluating 1D-CNN Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve.	147
Figure 6.16 The 1D-CNN model accurately predicts unseen data in five different categories.	151
Figure 6.17 Frame of video in class FD1 before and after applying face detection.	153
Figure 6.18 The proposed 1D-ConvLSTM model architecture	154
Figure 6.19 Layers of proposed 1D-ConvLSTM model.....	155
Figure 6.20 The Evaluation Metrics of 1D-ConvLSTM Model Performance.....	156
Figure 6.21 Loss, Mean Square Error, and Mean Absolute Error of 1D-ConvLSTM Model.....	157
Figure 6.22 The Confusion Matrix of 1D-ConvLSTM Model.....	158
Figure 6.23 The Evaluation Metrics of 1D-ConvLSTM Model.....	159

Figure 6.24 The Classification Report of 1D-ConvLSTM Model.....	160
Figure 6.25 Evaluating 1D-ConvLSTM Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve.....	161
Figure 6.26 The 1D-ConvLSTM model accurately predicts unseen data in five different categories.	165
Figure 6.27 Overview system (Transformer-based FER).....	166
Figure 6.28 The Evaluation Metrics of ViT Model Performance.....	167
Figure 6.29 Loss, Mean Square Error, and Mean Absolute Error of ViT model.	168
Figure 6.30 The Evaluation Metrics of ViT Model.	168
Figure 6.31 Evaluating Vit Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve	169
Figure 6.32 The Confusion Matrix of Vision Transformer Model.	169
Figure 6.33 The Precision-Recall Curve for Each DNN model.	172
Figure 6.34 Confusion Matrix for each proposed model.	173
Figure 6.35 Comparison between evaluated Models by Receiver Operating Characteristics Curve	175

List of Abbreviations

1D-CNN	1 Dimensional Convolution Neural
1D-ConvLSTM	1 Dimensional Convolution Long-Short Term Memory
AAM	Active Appearance Model
AAN	Artificial Neural Network
AFER	Automatic Facial Expression Recognition
AI	Artificial Intelligence
ASM	Active Shape Model
AU	Action Unit
AUC	Area Under a Curve
AUDN	AU-inspired deep networks
BERT	Bidirectional encoder representations from transformers
BiLSTM	Bidirectional Long Short-Term Memory
BP	Blood Pressure
CCU	Critical Care Unit
CGI	Computer-Generated Imagery
CK+	Extended Cohn-Kanade
CNN	Convolution Neural Network
DCNN	Deep Convolution Neural Network
DL	Deep Learning
DNN	Deep Neural Network
DRML	Deep Region and Multi-Label Learning
DT	Decision Tree
EEG	Electroencephalograms
ECG	Electrocardiogram
SEMG	Surface Electromyography
FACS	Facial Action Coding System
FD	Face Display
FE	Facial Expression
FER	Facial Expression Recognition
FN	False Negative
FP	False Positive
GANs	Generative adversarial networks

HCI	Human Computer Interface
HMI	Human Machine Interface
HoG	Histogram of Oriented Gradients
HR	Heart Rate
ICU	Intensive Care Units
IMED	Indonesian Mixed Emotion Dataset
JAFFE	Japanese Female Facial Expressions
KMU-FED	Keimyung University Facial Expression of Drivers
LBP	Local Binary Pattern
LLM	Large Language Models
MaFE	Macro Facial Expression
MiFE	Micro Facial Expression
MAP	Micro Action-Pattern
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
MTCNN	Multitask Cascade Neural Network
NEWS	National Early Warning Score
NLP	Natural Language Processing
NN	Neural Networks
PR Curve	Precision-Recall Curve
PRD-FE	Patients at Risk of Deterioration-Facial Expressions
R-CNN	Region-based Convolution Neural Network
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAM	Spatial Attention Module
SIFT	Scale Invariant Feature Transform
SFEW	Static Facial Expressions in the Wild
SwinT	Swin Vision Transformers
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

TPR

True Positive Rate

ViT

Vision Transformer

Chapter 1

Introduction

1.1 Background

An understanding of human feelings, behaviours, and intentions is based on interpreting their sentiments expressed by various cues, including verbal communication such as speech patterns and nonverbal communication such as body language, gestures, head nods, and facial expressions (FEs) (Manalu & Rifai, 2024).

In the healthcare field, communication is an essential factor for understanding patient health and well-being (Street et al., 2009). However, it can be difficult to communicate directly with patients in critical states for a number of reasons, such as unconsciousness, severe illness, inability to speak, under medication, cognitive impairment, mental disorder, or second language. In such cases, a medical team must rely on measurements like vital signs monitoring (including heart rate (HR), blood pressure (BP), temperature, oxygen saturation), pain assessment (using the visual analogue scale), imaging tests (including chest X-rays, CT-scans, ultrasound). Normally, facial expressions are not considered in these situations, and this is the main topic of this PhD thesis.

Deterioration is a critical state and a serious factor that may result in the death of the patient (Jones et al., 2013), and can impair patients' ability to communicate and convey their feelings, thoughts, and desires (Alasad & Ahmad, 2005). An early accurate clinical assessment known as a self-report is not always an option when communication with a patient is impaired due to some factors such as age, critical status, unconsciousness, language impairments, inability speak, and limited mobility because of critical illness and/or medications process (Herr et al., 2019; Alasad & Ahmad, 2005). Self-report is a costly procedure concerning human resources and difficult to carry out objectively. Equally, risk assessment by critical care nurses in intensive care units (ICU) can be non-objective and based on intuitive decisions (Odell et al., 2009; Madrigal-Garcia et al., 2018). In addition, patients admitted to ICU mostly face life-threatening illnesses and having no effective communication may threaten judicious decisions control. Therefore, early warning clinical signs of deterioration in health care are important indicators of the imperative to take

preventive measures for patient well-being and chance of survival (Ye et al., 2019). Consequently, a human-computer interaction (HCI) system that can perform objective, accurate measurements, and precise health assessments, would be advantageous for identifying early warning signs and prioritising patients in need of urgent medical care.

The face is an essential indicator of human feelings, a significant channel for transferring nonverbal communication, and a prime source of health information (Guo et al., 2023). Therefore, facial expression (FE) is an important marker for conveying our emotional state and intentions (Joseph & Geetha, 2020). Psychologists assert that FEs form 55% of daily human communication, far higher than verbal communication with speech (38%) or in writing (7%) (Mohan et al., 2012).

FEs are formed by the synchronised movement of facial muscles and play an essential role in the exchange of information and fostering social interactions (Albert Mehrabian, 1972). FEs reveal human sentiments, giving a glimpse into a person's state of mind through changes in key facial features. The study of (Ekman & Friesen, 1978) presented an innovative method called the Facial Action Coding System (FACS), which succeeded in characterising movements of facial muscles that surround the facial landmarks, namely the eyes, nose, and mouth into specific FEs. Seven universal facial expressions were identified, perceived as emotions of happiness, anger, contempt, disgust, surprise, fear, and sadness. Their study introduced a comprehensive set of anatomic non-overlapping facial muscle movements called action units (AUs) that are typically used to encode and taxonomize FE corresponding to a displayed emotion (Rudovic et al., 2015; Jeong & Ko, 2018). Each AU refers to the contraction of a particular group of face muscles, and FACS perceives the configuration of AUs comprising a FE. FACS can identify large numbers of emotions by identifying the set of muscular movements that comprise the FE.

Recently, various methods of facial expression recognition (FER) based on FACS have been developed, identifying the complex changes in facial movement that form universal FEs (Sato et al., 2019). There are large numbers of FEs due to the huge diversity in facial muscle movement, and AUs represent the features that need to be extracted to recognise a FE.

Combinations of facial AUs, and the relationships between them, result in a specific facial behaviour and FE. A set of AUs may comprise, for instance, the co-occurrent relationship of inner brow raiser (AU1) and outer brow raiser (AU2), or the mutually exclusive relationship of lip presser (AU24) and lips apart (AU25).

In the medical field, observing and evaluating FEs by healthcare professionals can provide a valuable insight into a patients' feelings without causing them discomfort (Kuramoto et al., 2019). However, recognising the various types of FE using self-reporting is a challenge. Firstly, FE evaluation depends on the intuition and experience of healthcare professionals and psychologists who need to recognise and interpret interactive changes conveyed through the FE (Kuramoto et al., 2019). Furthermore, patient monitoring is based on observation by nursing staff, and such measurements that indicate patient's health deterioration may not be noticed in the time between observations. Moreover, the time and effort spent on training to statistically score the AUs is costly (Casella et al., 2023; Nagireddi et al., 2022).

Other applications of facial analysis have been designed in connection with age estimation, gender classification, face detection, face recognition, face posing and expression, and blink detection. Continued progress in facial analysis demonstrates its usefulness and effectiveness, as reflected in its increasing application in various disciplines. For instance, computer-aided FER is an interesting field offering various techniques to interpret human emotions from FEs. Together with effective human-machine interaction (HMI) interfaces, these technologies have significant potential to recognise and understand emotions and intentions.

There has been a significant growth in the use of digital images, machine learning (ML), and deep learning (DL) in facial recognition and human-computer interaction due to the availability and cost-effectiveness of high-end cameras. Facial image detection, analysis, and recognition have achieved remarkable real-time results.

However, identifying the intensity of facial AU is still a challenging task due to factors like head position, illumination, age, subtle expression changes, or the involvement of other specific sets of related and unrelated AUs (Rudovic et al., 2015). The main motivation for using ML and DL algorithms is their high performance in detection and recognition tasks. Robust artificial neural network (ANN) models can be designed and trained on large and diverse datasets to satisfy the given requirements of automatic recognition and classification. (Hardas & Pokle, 2017). The automatic recognition of deterioration in health is essential in the healthcare domain as a supportive factor for patient recovery in ICUs, and for admission to critical units after surgery. Precise deterioration assessment from facial expressions could be highly beneficial as an automatic early warning system, facilitating risk assessment earlier than statistical methods.

This thesis presents automatic facial expression recognition (AFER) as a technique to support healthcare professionals by providing specific information on health status without the need for previous knowledge or special skills. Various approaches exist for feature extraction from the fine details of a dataset. To capture the various aspects of facial data using ML and DL, this thesis proposes a combination of appropriate feature extraction approaches for a comprehensive, sensitive, precise, and robust analysis of a patient's health status. Guided by the work of Madrigal and her colleagues, this thesis investigates and proposes five models based on different pre-processing, feature extraction, and AI methods, to predict health conditions by recognising early signs of deterioration. The study aims to recognise facial expressions with high levels of accuracy and generate realistic datasets that mimic the exact expressions of patients whose health shows signs of deterioration within a given time frame. Due to the difficulty of obtaining real-data samples from patients in critical care units, the generated datasets will make a significant contribution to knowledge. The highlighted models are as follows.

- The first model introduces one of the most common deep neural network models known as convolution neural networks (CNN) to learn and predict facial features from generated characters created by special software such as Blender and the first order motion model (FOMM).
- The second model focuses on convolution neural networks and long short-term memory (LSTM) to replicate predictions from the detection and recognition of facial expressions from a set of action units.
- The third model is based on feature extraction and the random forest (RF) as a classifier with high quality prediction ability.
- The fourth model is based on feature extraction and the support vector machine (SVM) as a classifier, yielding high-quality predictions.
- The final model involves the transformer deep learning algorithm used in large language models (LLM) to determine whether it can produce robust predictions from minute facial expressions.

The work reported here concerns the development of an automatic facial AU recognition system capable of detecting patient health deterioration in critical care wards based solely on FEs. The emotional states expressed are detected in real-time using fully automated computer algorithms whose input is facial images acquired from any imaging device, such as a webcam.

1.2 Related Work

Human facial expressions are powerful nonverbal cues of emotions, state of mind, and health, and FER is essential in human communication and interaction (Jaswanth & David, 2020). In 1971, one of the earliest works on facial expression was presented by Ekman and his colleagues (Ekman & Friesen, 1971), who developed a theory of universal FEs for specific emotions by observing films of social interactions in different cultures.

Seven years later, the facial action coding system (FACS) (Ekman & Friesen, 1978) was presented, identifying AUs that represent facial muscular movements. In 1995, Gosselin et al. (Gosselin et al., 1995) described an experiment that included six participants from Canada to present emotions based on scenarios corresponding to six types of FEs. FACS shows that certain AUs represent specific FE, like AU 6 and AU 12, that appear frequently in a happy expression, or AU9 that is rarely observed in a disgusted expression. (Sato et al., 2019). Furthermore, several non-predicted AUs were observed frequently in most FEs. Later, the authors (Scherer & Ellgring, 2007) asked 12 professional actors in Germany to present facial expressions corresponding to various ranges of emotions. According to the FACS analyses, the outcomes of the experiment did not prove the existence of many theoretically predicted AUs for basic and non-basic emotions. Therefore, in recent decades, computer-aided vision has been identified as an essential tool for healthcare professionals. In 2011, Lucey et al. (Lucey et al., 2011) built a UNBC-McMaster database containing 200 video streams taken from 25 patients suffering from shoulder pain. Video frames were labelled based on the work of Prkachin and Solomon (Prkachin & Solomon, 2008) and the FACS metric presented by Ekman, Friesen, and Hager, which codes different facial muscle movements with various intensity levels (Ekman et al., 2002). The UNBC-McMaster pain expression archive database has been used to propose new models for FE pain detection. Lucey et al. (Lucey et al., 2011) published baseline results with the dataset that used support vector machines (SVM/AAM) system to extract facial landmark features to predict pain action units (AUs) and the PSPI to score the presence of pain.

In these and similar studies, facial AUs have been typically used to encode facial activity corresponding to different facial expressions such as pain or anger. In 2015, Rudovic et al. (Rudovic et al., 2015) described the difficulty of reliable AU intensity estimation in differing contexts, such as light intensity, head pose, or variability in FE. As a result, the analysis of facial expressions has received a lot of attention. In 2013, an investigation study using FACS proposed by Gross and his colleagues (Gross et al., 2013) uncovered that health

professionals usually recognise sadness and fear expressions in patients at risk of deterioration. In 2018, (Madrigal-Garcia et al., 2018) a collaboration between North Middlesex University Hospital, University College of London Hospital, and the GMPR Research Group at Sheffield Hallam University proved that patterns of facial AUs can be used as predictors of admission to critical care. The study analysed AUs in both the upper and lower regions of the face, including head tilting direction, and the National Early Warning Score (NEWS) was used to collect clinical metrics. This seminal work was the starting point for this PhD thesis. While that work used a trained psychologist to recognize facial expressions, the aim of this thesis is to develop an automatic facial expression recognition (AFER) based on the work of Madrigal-Garcia and colleagues.

In recent years, AFER has become a crucial part of various human-computer interaction applications (Li et al., 2013; Chen et al., 2019). Considered a multidisciplinary research field, it has been applied in computer vision, machine learning, psychology, neuroscience, and cognitive science (Gunes & Hung, 2016). In 2016, Jaiswal and Valstar (Jaiswal & Valstar, 2016) presented a combination of Convolution Neural Networks (CNN) and Bi-directional Long Short-Term Memory Networks (BiLSTM) that can detect facial AUs. In 2017, Sang et al. (Sang et al., 2017) introduced CNN model that can recognise facial emotions, in which the output layer includes seven neurons that are labelled according to the seven universal FEs. The purpose is to classify each image as one of the universal facial expressions. One year later, Chen et al. (Chen et al., 2017) presented a CNN that uses a convolution kernel for feature extraction and a max pooling operation to minimise the dimensions of the extracted features. This AFER system also classifies each facial image as one of the seven universal FEs. In an early AFER investigation by (Al Taei & Jasim, 2020) using the Japanese Female Facial Expressions (JAFFE) database, a CNN trained with different grey-scale images was able to classify FEs as one of the seven universal emotions with 100% accuracy. The work of (Mohan et al., 2021) introduced deep convolution neural networks (DCNN) for recognizing facial expressions. The researchers proposed a 2-stage approach based on five databases. The first stage focuses on identifying geometric local features in the human face using a gravitational force descriptor. In the second, the descriptor is fed into the DCNN model, exploring the holistic features. In summary, facial AUs have been employed to encode different facial expressions with varying degrees of success and with intrinsic model limitations.

1.3 The Significance of this Research

Facial expressions are vital identifiers of human sentiments and non-verbal cues which uncover our emotions. The early signs of deterioration in health that are revealed by facial expressions can be automatically detected and recognized using appropriate methods.

This thesis presents an automated medical examination and risk alerting process by acquiring information from face segments of interest within facial images to derive precise real-time data on the stability of a patient's medical condition. Machine learning algorithms are selected in the design of a prediction or decision-making model based on their ability to transform raw data into valuable insights.

The suggested method concentrates on investigation of the main facial features that can help a CNN or LSTM recognize patterns of deterioration in patients in critical care units, ignoring unnecessary and misleading features that confuse deep learning networks through the training process. The project focuses on five essential classes of facial expression that show deterioration reported in the work of Madrigal-Garcia and her colleagues (Madrigal-Garcia et al., 2018).

Detecting early signs of health deterioration has a significant impact on human survival and improving their well-being by providing healthcare professionals with alerting signals for appropriate immediate support actions. Specific FEs indicating deteriorating health have been proposed by researchers from NHS at Middlesex Hospital and Sheffield Hallam University who observed and recorded characteristic FEs among patients about to be admitted to critical care units (CCUs).

This thesis seeks to design models to acquire information from images of specific regions of the face, including upper and lower parts, facilitating computer-aided facial analysis based on facial muscle movements. Using computer vision techniques, the thesis aims to facilitate automated medical assessment, especially in situations challenging to observe like deterioration trends over many hours.

Therefore, this project provides a systematic review to assist the healthcare professional, presenting analysis of methods that produce useful information on the health status of patients through the reading and analysis of micro-expression intensity. The main driver for developing techniques to detect deterioration trends from facial muscles is to overcome existing limitations of cost and subjectivity. Such an early warning system has high clinical value.

1.4 Motivation

The exchange of information through FEs during social interaction or medical assessment is a substantial contributor to understanding an individual's emotional state, especially when verbal communication is impaired. Several facial attributes such as facial landmarks, action units, and micro and macro facial expressions can be employed to derive and analyse significant features of facial expressions.

Various AFER studies have yielded high accurate results in processing and interpreting images and video frames, including identifying and tracking facial attributes. With a crucial role in FER, deep learning algorithms trained on large datasets of FEs can recognize patterns and correlate the intensity of action units with specific emotions. Algorithms such as CNN and LSTM can derive and categorise attributes from the changes in facial muscle movement.

Therefore, this thesis develops AFER methods to analyse facial expressions reflecting trends in a patient's health by designing an early warning system based on deep learning techniques with the following characteristics.

1. A monitoring and alert system will send signals and information showing patient status in real-time to significantly facilitate decision-making.
2. Involuntary micro-expressions that appear and leak over the face expressing allow the real emotional state to be detected without any control or manipulation from the patient, not only benefiting the healthcare sector but other disciplines such as crime investigation, social interaction, addressing security issues, and education.
3. The ability to recognise subtle FE cues of patient health state will have a significant impact on assessments such as pain assessment, autism, and depression.
4. Building interaction between healthcare professionals and patients through effective interpretation of facial expressions can improve survival rates and build an empathetic healthcare environment.

With evidence from recent research recording high accuracy in AFER using machine learning and deep learning techniques, this thesis advances current knowledge of automated techniques to identify patients at risk of deterioration.

1.5 Problem Statement

Facial expressions are the innate outcome of voluntary and involuntary emotional responses, and their accurate measurement and interpretation is a challenging task even for trained professionals. Therefore, employing machine learning and deep learning techniques, especially CNN and LSTM algorithms, to extract, describe, and classify patterns of micro and macro facial expressions can significantly impact accurate understanding of emotional states.

Deep learning-based FER methods face the challenge of scarcity of databases of FEs. The core problem that this project faced was generating a realistic dataset of FER for patients in deteriorating health that closely mimics real data.

In addition, deep learning models often struggle with capturing and extracting nuanced and subtle changes in muscle movement. The preparation phase, when implementing some of the extraction methods, faced challenges like the intensity of micro facial expressions, the imbalanced distribution of data per class, and varying frame rates of the captured facial images and videos. Furthermore, the data are virtual not real, potentially resulting in unrealistic datasets. Therefore, employing recent techniques to extract as much realistic data as possible is crucial. The performance of machine learning models can suffer in the real world due to unmeasured circumstances, and unaccounted environmental factors.

1.6 Aim and Objectives

The aim of this project is to develop and demonstrate an AFER system that can provide continuous, real-time monitoring of patients' facial expressions, which is especially useful in high-dependency environments like intensive care units (ICUs), where patients may be non-verbal due to their critical state or sedation. AFER offers objective and consistent monitoring without the subjectivity or fatigue that can affect healthcare professionals' observers, and providing early signs detection help for timely intervention, which can significantly improve patient well-being and potentially reduce recovery time.

The proposed AFER framework is based on machine learning (ML) and Deep Learning (DL) algorithms for allocating micro and macro facial expressions, extracting features, and categorizing FEs. The following objectives are identified:

1. **Development of an AFER Framework:** To design and implement an AFER system using machine learning (ML) and deep learning (DL) techniques that can accurately identify and classify micro and macro facial expressions, enabling the detection of early signs of patient health deterioration in high-dependency settings.
2. **Creation of a Synthetic Facial Expression Dataset:** To generate a more realistic and representative dataset for training AFER models by developing avatars that express specific facial expressions and transferring these to images of real faces, enhancing the system's accuracy in detecting health-related facial expressions.
3. **Improvement of Feature Extraction and Recognition Methods:** To investigate and optimize feature extraction methods to recognize facial expression patterns, focusing on those linked to health deterioration, while addressing challenges like variations in face poses, lighting, and facial landmarks.
4. **Advanced Deep Learning Model Design:** To design and develop state-of-the-art deep learning models, including Convolutional Neural Networks (CNN), ConvLSTM, and Transformers, for hierarchical feature learning, real-time processing, and efficient facial expression analysis without the need for manual feature engineering.
5. **Application of AFER in Real-World Clinical Scenarios:** To evaluate the system's robustness in real-world conditions, focusing on timely and accurate detection of patient health deterioration, especially in critical care and general hospital environments. The AFER system will act as a preventive measure, aiding in health assessments and sending alerts to healthcare professionals for timely intervention.
6. **Impact on Healthcare Outcomes:** To demonstrate the potential of AFER in reducing critical care admissions, preventing disease progression, and lowering healthcare costs by facilitating early detection of health issues, reducing the duration of hospital stays, and optimizing medical interventions.

This thesis will contribute to improving healthcare by designing and developing an early detecting automatic warning system that detects the early signs of deterioration through facial expressions. The presented FER model is designed and adapted for involuntary facial micro-expressions which is considered a challenging task due to their brief appearance, very short duration, low intensity, and limited pattern changes.

1.7 Expected Contributions to Knowledge

This thesis presents the state-of-the-art results on generating and modelling synthetic database and automated deterioration prediction through FEs. The expected contributions to knowledge from this thesis are highlighted as follows.

1. **Development of a Real-Time, Objective Monitoring System:** This thesis presents an automatic method with state-of-the-art performance that could have the ability to detect subtle facial cues related to health deterioration to improve real-time patient monitoring providing continuous, and objective monitoring of patients' FEs which has significant impact on admission to high-dependency environment such as ICUs and Critical Care Units (CCUs).
2. **Enhanced Communication for Non-Verbal Patients:** The AFER system provides innovative methods to improve communication through identifying non-verbal cues like FEs that indicate a patient's critical condition especially for those who cannot communicate verbally due to sedation, or cognitive impairments such as stroke, comma, dementia, etc.
3. **Improving Timely Clinical Interventions:** The proposed AFER system has the potential to trigger early warning signs of deterioration for enables immediate clinical interventions to act before reaching critical stages and life-threatening which can result in increasing the chances of patient survival and recovery rates.
4. **Generating and Developing a Synthetic Facial Expression Dataset:** A novel synthetic dataset of macro and micro facial expressions was developed, reflecting both stable and deteriorating health conditions. This dataset simulates real-life scenarios by transferring facial expressions from avatars to human faces, making it a valuable resource for training and testing AFER models. It enhances the system's ability to accurately identify expressions under diverse conditions, improving the generalizability and robustness of the models.
5. **Development of Two State-of-the-Art Deep Learning Models:** This thesis adapted two high-performing deep learning models: the 1D-CNN combined with ConvLSTM, and a second model utilizing a Convolutional Neural Network (CNN). Both models achieved state-of-the-art results in predicting patient health deterioration. The 1D-ConvLSTM model achieves an accuracy of 99.89%, making it highly effective for real-

time monitoring (see chapter 6 and published paper Al-Tekreeti et al., 2024). These models can be integrated into future healthcare applications, reducing reliance on self-reporting methods and enabling more efficient patient monitoring.

6. **Impact on Healthcare Outcomes and Cost Reduction:** The AFER system's ability to detect early signs of health deterioration contributes to improving healthcare outcomes by facilitating early interventions. This not only reduces the need for critical care admissions but also prevents the progression of diseases and minimizes the need for extensive treatment. As a result, the system can significantly lower healthcare costs by reducing the duration of hospital stays and optimizing medical interventions, leading to better patient management and resource allocation.

Overall, by recognizing facial expressions that indicate health deterioration, this research enhances the ability to provide timely, empathetic, and effective patient care. The contributions made in this thesis pave the way for the development of advanced healthcare monitoring systems that can lead to improved recovery rates, reduced healthcare costs, and human well-being.

1.8 Publication

- Al-Tekreeti, Z., Moreno-Cuesta, J., Madrigal Garcia, M. I., & Rodrigues, M. A. (2024, August). AI-Based Visual Early Warning System. In *Informatics* (Vol. 11, No. 3, p. 59). MDPI.

1.9 Methodology

Development of the AFER system involves several stages, starting with generating a synthetic dataset and ending with classifying the FEs. The overall methodology is the waterfall method with continuous feedback review and improvements. Figure 1.1 depicts a waterfall model that illustrates the essential stages of the proposed project at a high level of definition.

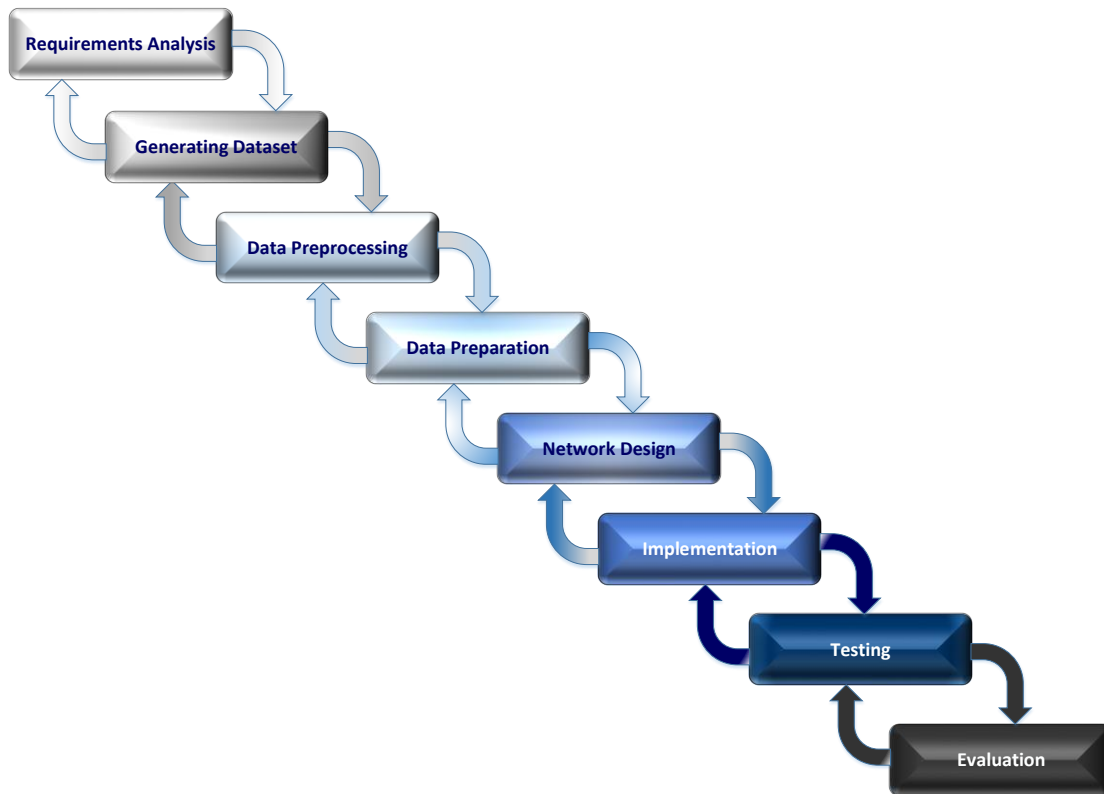


Figure 1.1 The waterfall model with feedback.

In what follows, specific steps are highlighted and customized to this project, which involves data generation and deep learning methods. Therefore, the following steps comprise the proposed methodology.

1. Dataset generation

- a. *Define dataset objectives and requirements.* The dataset must allow recognition of five types of FEs that indicate patient deterioration risk. Data requirements include types of facial expressions (identifying AUs that form each type of proposed FE), diversity of participants (age, gender, skin tone, ethnicity), and diverse environmental conditions (lighting, background).
- b. *Design avatars.* Design or select base models for avatars using 3D modelling software (e.g., Blender, Maya, FacsHuman), and customize their features to ensure diversity in facial features and expressions (age, gender, skin tone, ethnicity).
- c. *Animate expressions.* Animate the predetermined expressions using keyframe animation or motion capture technology to create smooth transitions and realistic

actions. Then, review animations to ensure that each smoothly and accurately conveys the intended FE and make necessary adjustments based on feedback.

- d. *Data annotation.* Label each animation frame or video with the corresponding facial expression and add metadata for each frame or sequence, detailing the FE type and avatar characteristics.
- e. *Rendering.* Render animations to ensure consistency in the appearance and combination of AUs, creating high-quality images or video frames with appropriate frame rate (20, 25, 30, or 60 per second).

1. Dataset preprocessing

- a. *Face detection.* Apply face detection algorithms to crop images to only include the face, removing unnecessary background. Locating and isolating faces from irrelevant and undesirable data can promote performance, robustness, and accuracy of an AFER model.
- b. *Normalization.* Standardize image sizes and colour schemes for uniformity to adjust the range of pixel intensity values, typically 0 to 255. This is particularly important when images are obtained under various lighting conditions.
- c. *Augmentation.* Apply data augmentation techniques such as rotation, flipping, and scaling to expand the dataset size, enhance the quality of the dataset, and improve model robustness. For the proposed model, the data augmentation techniques are changes in lighting, and slight modifications to expressions.
- d. *Oversampling.* After splitting data to train and test the dataset, employing oversampling techniques for training the dataset can mitigate the drawbacks of imbalanced classes, where one class or category has significantly more examples than another. This imbalance can lead to biased machine learning models that perform poorly on the minority class. Oversampling helps address this issue by increasing the number of instances of the minority class.

2. Feature extraction

- a. *Manual feature extraction.* Identify key facial landmarks (eyes, eyebrows, mouth, nose) that are essential for recognizing expressions.
- b. *Automated feature extraction.* Employ deep learning models, particularly convolution neural networks (CNNs), to automatically extract relevant features from the facial data.

3. Model Training

- a. *Select appropriate model.* Selection of an appropriate model (e.g., CNN, LSTM, or hybrid models) is based on dataset type and task objective.
- b. *Model design.* Determine number of layers and identify type of each layer.
- c. *Configure hyperparameters.* Set the learning rate, number of layers, batch size, etc.
- d. *Train model.* Use the labelled dataset to train the model to recognise and classify facial expressions.

4. Model Evaluation

- a. *Validation.* Test the model on a separate validation dataset to evaluate its accuracy, precision, recall, and F1-score.
- b. *Tuning.* Adjust model parameters based on performance metrics to optimise accuracy and reduce overfitting.

5. Deployment (future work)

Real-time AFER system conducted on real-world data samples.

6. Monitoring and updating

Continuously monitor the system performance and adjust as needed. Evaluation metrics provide feedback to the designer and developer with information related to model performance, identifying refining requirements.

7. Model retraining

Periodically retrain the model with new data to adapt to changes in expression representation or to improve accuracy.

Figure 1.2 illustrates the main steps of the proposed methodology.

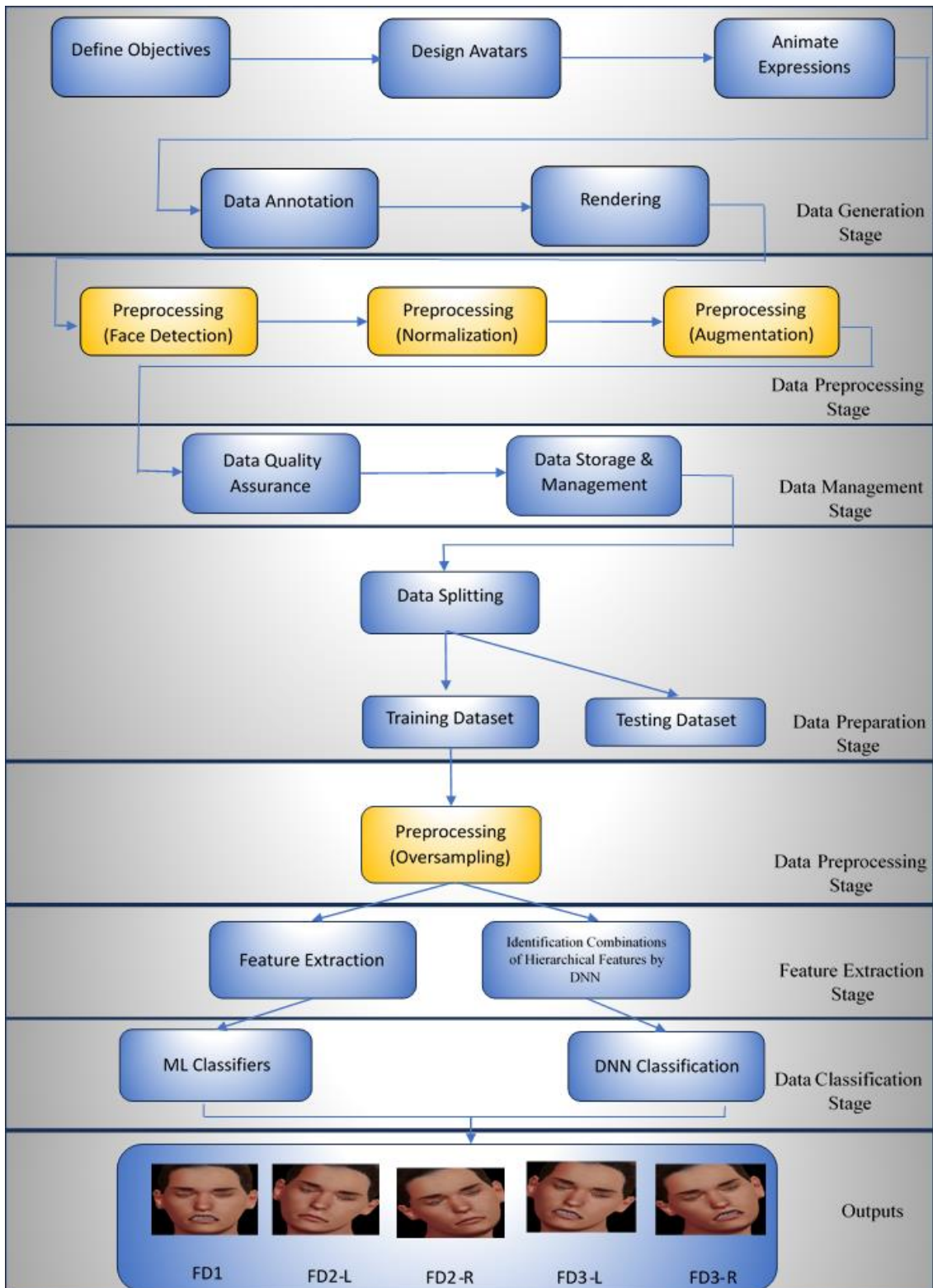


Figure 1.2 Methodology of Proposed AFER system.

Figure 1.3 depicts virtual simulation environment for Intelligent ICU includes camera, video monitoring, and AI-VIEWS provides patient status updates by analysing the patient's facial expressions in real time.



Figure 1.3 Virtual Simulation Environment depicts Intelligent ICU includes the proposed AI-VIEW system.

This 3D virtual environment shows an ICU that includes a virtual patient (avatar) lying on a hospital bed, and a healthcare professional stands beside him, observing the patient's status through monitoring equipment. In addition, this virtual ICU includes a camera to record patients faces and transmit the data to the Artificial Intelligence-Virtual Early Warning System (AI-VIEW) as an AFER model to analyse and recognise patterns of FEs in real time. Furthermore, the monitors, which display various views and close-ups of the patient's face, along with facial landmarks used in this project for face detection and AI-VIEWS prediction, indicate whether the patient is at risk of deterioration.

1.10 Thesis Structure

The remainder of this thesis is organized as follows.

- Chapter 2 provides a review of the literature relating to facial expression recognition techniques. It reviews the concept of facial expression and the background of the techniques employed for FER. It presents the key components of FER systems with reference to previous work. Furthermore, the obstacles faced in generating and preparing datasets are discussed and solutions are presented. Applications of FER in various disciplines and research areas are highlighted.
- Chapter 3 introduces the theoretical background of the deep learning techniques employed in this thesis. It explains the selection criteria for these methods and their impact on system performance.
- Chapter 4 highlights the dataset creation methodology and the methods employed in analysing, generating, and preparing the dataset. It describes solutions for creating a realistic dataset that best mimics real-world data.
- Chapter 5 describes feature extraction methods and ML classifiers, presenting the pipeline of FE recognition and classification using facial images and videos.
- Chapter 6 describes the proposed designed DNN models, presenting the pipeline to implement FE recognition and classification using facial images and videos.
- Chapter 7 concludes this thesis with a summary of the main findings, which highlights the contributions that the research makes to the healthcare field. It also highlights the limitations, recommendations based on the findings, and potential avenues for future work.

Chapter 2

Facial Expression: A Literature Review

2.1 Introduction

In the billions of human faces that vary in shape and landmarks due to different cultures and ethnicities, facial expressions are the universal means to express emotional states. Over recent decades, the world has witnessed a revolution in the recognition of human emotions using verbal and non-verbal communication, especially for health and well-being assessment. Communication and transferring information without the use of words through non-verbal cues is part of human interaction, and correct interpretation of these cues is central to successful communication.

Over the past few decades, researchers in various disciplines, particularly social science, psychology, medical science, and technological science, have studied the interpretation of non-verbal cues such as facial expression, eye contact, posture, gestures, touch, and proximity, among others. In non-verbal communication, facial expressions are highly informative of human emotions. Fundamentally, facial expressions occur due to the contraction, relaxation and motion changes of facial muscles beneath the skin in response to an individual's emotional state and can convey the complete spectrum of sentiments. Each group of muscle movements is associated with a specific emotion that forms a particular expression.

Activation of facial muscles can form wrinkles, lines, and folds and alter facial landmark positions. The richest attributes of facial expressions appear in the upper and lower segments of the face, especially in the areas surrounding the eyes and mouth.

For instance, blinking eyes are considered an indicator of whether a person is lying or nervous. Smiling is an indicator of happiness while a frown indicates disgust or sadness. This sign language is considered a basic communication channel between people, regardless of their cultural backgrounds.

Action units are components of a facial expression, and each AU has a particular position in the neutral face. When facial muscles are activated by the limbic system to contract or relax, a specific set of AUs assumes new positions to provide a specific facial expression.

For example, an expression of surprise can be decoded by a combination of raising inner and outer eyebrows, raising the upper lid, and opening the eyes and mouth widely (AU set 1, 2, 5, 25, and 26). Hence, each set codes for a different expression.

The limbic system is responsible for processing human feelings, and emotion signals are translated through the facial nerve to activate facial muscles. The facial nerve carries the nerve fibres and is the motor of facial muscle movement and facial expressions. Figure 2.1 illustrates the facial muscles and facial regions of interest in FER. Figure 2.2 highlights with red circles facial muscles that form AUs for patients at risk of deterioration.

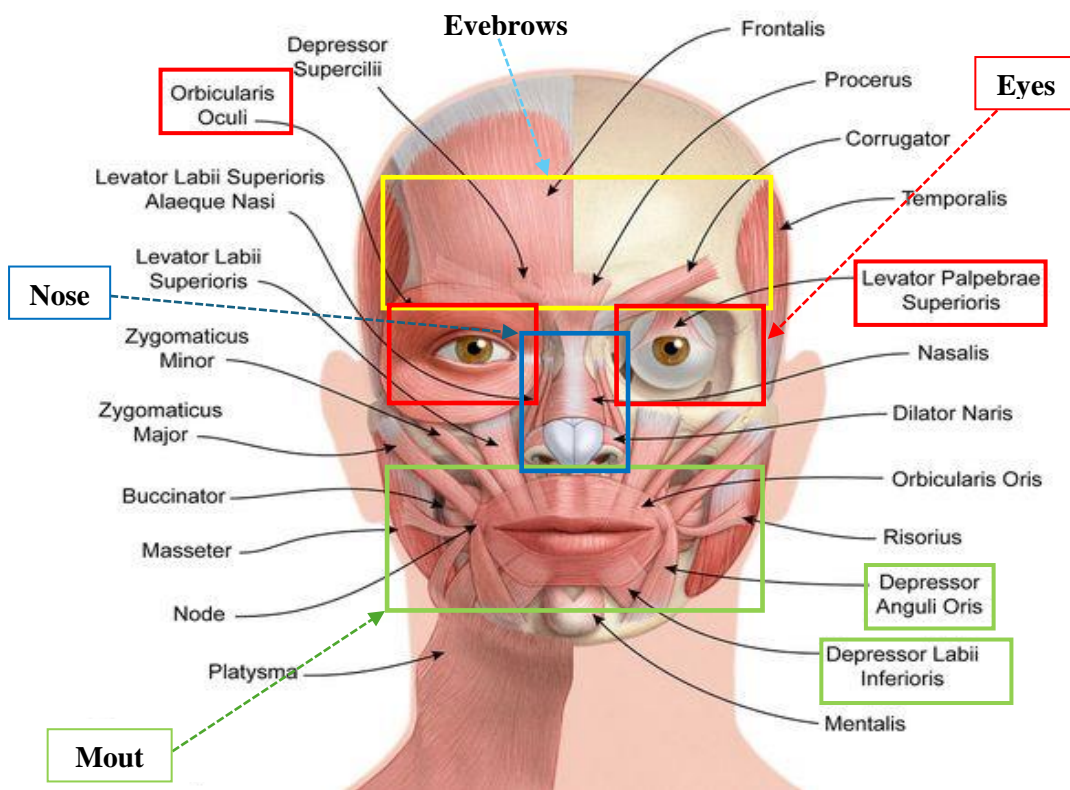


Figure 2.1 Main facial expression muscles of face and key facial areas division (Zhao et al., 2023).

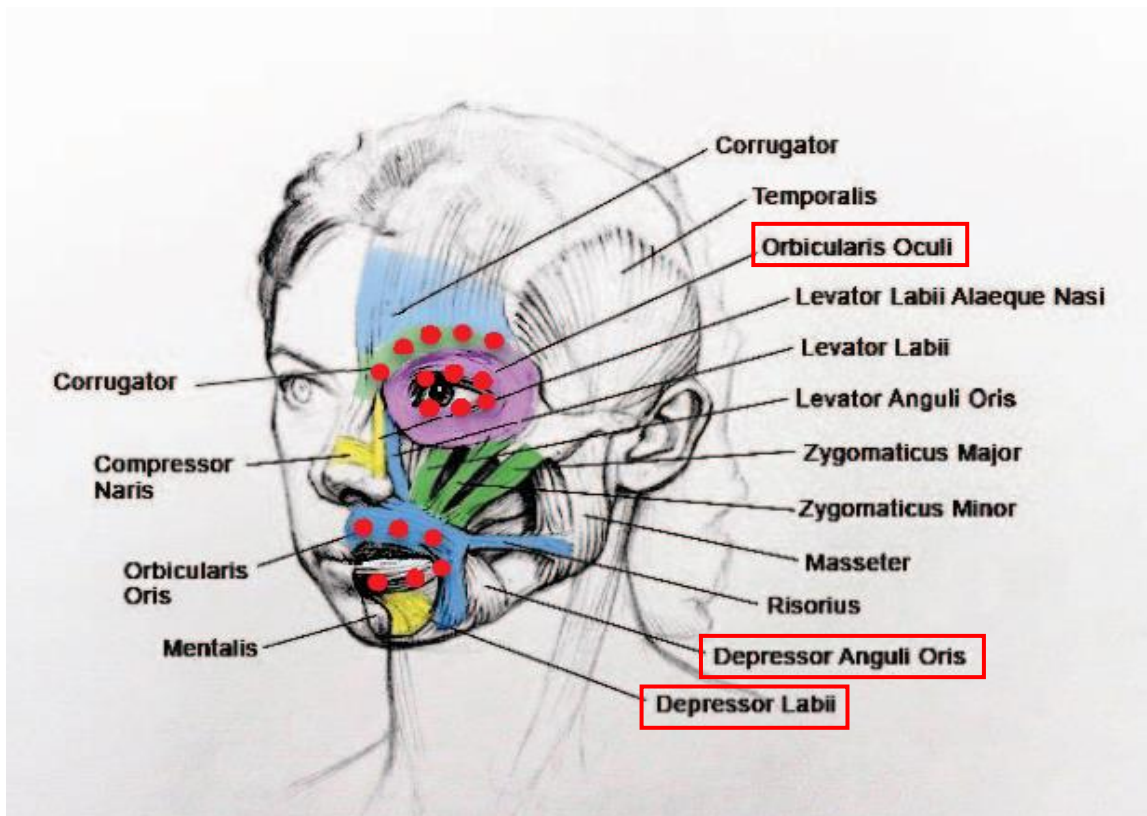


Figure 2.2 Main facial expression muscles of (Fehrenbach & Herring, 2015) Areas and facial muscles that reflect deterioration critical condition of patients are spotted with red circles.

In the healthcare field, communication is essential for understanding a patient’s well-being, but can be impaired or impossible for many reasons. Therefore, interpreting nonverbal cues is imperative because they are generated directly by the limbic system. Designing FER systems capable of interpreting and decoding the information in facial expressions has significant potential for helping healthcare professionals understand a patient’s psychological and emotional status. FER systems can be designed using manual and automatic approaches. Statistical methods require highly trained observers, often psychologists, to visually recognise and assess the facial cues. This method has some key drawbacks. Differing observer experience and background can result in inconsistent patient status assessments. The method is time-consuming and demands extensive effort from observers because some health conditions require prolonged observation to capture changes in specific facial expressions. Moreover, the capture of nuances and rapid changes in facial expression is a significant challenge.

All these limitations have been mitigated by automated FER. The average accuracy rate achieved by professional observers is around 50%, while automatic FER systems have achieved much higher degrees of accuracy. Automatic methods are quicker and less costly and are also less sensitive to environmental conditions due to their ability to eliminate noise and unwanted data.

Though dynamic machine learning methods, particularly deep learning algorithms, have shown remarkable results compared with the usual static methods, they still face obstacles and drawbacks. The model's performance depends heavily on the quality and quantity of the training data. Deep learning models require large amounts of labelled data in the training stage. Deep learning models are sensitive to noise and variations of light, pose, and other external and internal environment factors and can also suffer from overfitting, capturing noise, or unwanted features, resulting in the generation of low-quality data.

However, the scarcity of databases is considered the major obstacle to automated methods, so generating an adequate and qualified dataset is essential. The quality and reliability of model performance depend on providing a database that includes a wide range of data suitable for achieving the objective and effective across the various populations, cultures, and environments in real-world conditions. For this project, the database includes diverse data from different cultures, ages, and skin tones. Addressing these challenges and obstacles before and during the model design is essential for reliable performance.

This chapter includes a systematic review of facial expressions and their types, including micro and macro facial expressions and recognition methods related to this project. The subsequent sections detail micro- and macro-expressions and their applications in different disciplines. Section 2.4 then describes the generated database which is employed for evaluating the performance and accuracy of the designed models. Finally, the chapter ends by presenting the core components of the presented automatic models.

2.2 Facial Expressions (FEs)

Innate facial expressions are a source of emotional information that play a vital role in nonverbal communication facilitating connection across societies. FEs can be categorized into two kinds: micro-expressions and macro-expressions.

Recent research on FER using macro expression as the main source of information has achieved significant accuracy. However, while some researchers have recorded exceptional

results, the study of micro-expressions is still in progress. Major factors that distinguish them are the intensity, duration, and ability to control these different FE types. Macro FEs are prominent and can be easily recognized and detected with high accuracy since they last from 0.5 to 4 seconds. Micro FEs typically last less than 0.5 seconds and can be characterised as subtle, brief, and involuntary. Research into micro FEs therefore began later and has been a challenging task. Macro expressions are voluntary, meaning they can be controlled and manipulated by an individual, and therefore cannot be a legitimate measurement tool for identifying emotional cues. However, involuntary micro FEs cannot be controlled by an individual and attempts to consciously repress these cues will fail. The focus of this project is to present an effective AFER that can capture macro and micro facial expressions of patients to contribute to healthcare assessments by determining the need for admission to critical care units, extra care, or discharge.

2.2.1 Micro-Expressions

Facial micro-expressions (FMiE) convey unconscious emotions and have multiple uses in diagnosing medical conditions and enhancing national security.

(Li et al., 2022). The extrapyramidal motor system, encompassing the subcortical nuclei, plays a central role in maintaining posture and regulating involuntary motor function, including FMiE. Genuine emotions are not associated with the cortex and their occurrence is involuntary (Rinn, 1984). FMiE can be described as subtle manifest, involuntary, low intensity, rapid movement, and most authentic expressions which are typically exhibited for a very brief limited time to provide a deeper dive into a human's genuine emotion (Dewmini et al., 2021). Therefore, recognizing and distinguishing these expressions is a challenging task for the untrained eye.

By revealing authentic, genuine emotions that an individual may try to suppress or conceal, FMiE can provide researchers with a deeper knowledge of the complexities of no-verbal communication, state of mind, and attitude. This information is of significant value in diverse fields such as social science, psychology, crime investigations, clinical diagnosis, law enforcement, and artificial intelligence science that are interested in understanding true emotions, behaviours, and intentions. As a response of the limbic system, the contraction and relaxation of facial muscles in a certain order reveals FEs. FMiE can give researchers valuable insights into the limbic system to understand sentiments.

FMiE will find important applications in law enforcement and security by revealing true intentions before they can be concealed by controlled macro expressions (Ekman, 2009) . The ability to detect fake expressions will have an impact on lie detection in crime investigations (Ekman & O'sullivan, 1991). In the medical area, a patient's psychological status is considered an essential factor in understanding their sentiments, and FMiE is a source of authentic information for healthcare professionals (Yan et al., 2013). This can be vital for appropriate care decisions, especially when verbal communication is impaired or for detecting signs of autism, schizophrenia, and mental disease. Hence, FMiE can contribute to the assessment of various medical conditions.

The main characteristics of Facial Micro-Expressions are listed as follows.

1. Fleeting. The period of appearance is significantly shorter than macro expression.
2. Involuntary. Micro-expressions naturally occur without human interference, cannot be concealed or repressed, and so reveal genuine emotions. Conversely, macro expressions are voluntary and conscious actions.
3. Subtle. Micro expressions are fine and precise, resulting from rapid muscle movement. They can easily go unnoticed and can be obscured by longer-lasting expressions.
4. Authentic. Because they are involuntary, they express genuine human feelings.
5. Low intensity. Micro expressions leave low intensity imprints compared with macro expressions.

Micro expressions can be observed and determined manually by highly trained psychologists, or automatically by machine learning. The accuracy of experiments using manual statistical methods has been recorded at around 50%, while the automatic methods have achieved an average recognition rate of at least 75%.

Genuine spontaneous micro-expressions occur as a response to authentic true emotions (Yan et al., 2014). The characteristics of these expressions are subtle, fine, brief, fleeting, and lasting for a brief duration of time (Ekman & Friesen, 1969). Various facial muscles around the eyes, eyebrows, mouth, and other facial segments are activated by brain signals in response to an emotional stimulus to produce micro expressions. The following factors contribute to the appearance of this kind of expression.

- Facial muscle can be triggered by emotional stimuli factors or cognitive processes, some of them unexpected sudden events, thoughts, and memories that result in the occurrence of an immediate and true emotional response (Ekman, 1992).

- The limbic system is mostly formed by regions of the brain such as the amygdala and the insula which send signals to make rapid and spontaneous activation of facial muscle, leading to this kind of precise brief expressions. Even when some people try to stifle or suppress their original emotions there is still a momentaneous leak that reveals a genuine emotional state through micro-expressions (LeDoux, 2000).

It occurs before the individuals try to conceal their genuine emotional state which may not align with their true intentions or appeared expressions. Identifying and interpreting these brief cues before the individuals have a chance to conceal them can provide valuable deep knowledge into a human emotional state and aid in analysing and understanding their hidden intentions (Ekman, 2009).

In the discipline of psychology and social science, detecting and identifying these subtle expressions accurately requires significant training and experience in nonverbal communication due to their fine exhibition that lasts for a fraction of a second.

- For the micro-expressions, the contraction of facial muscles occurs through an extremely short span, and for that reason, the contractions and relaxation are almost imperceptible to the untrained eye (Ekman, 2009).
- Because of the characteristics of MiFE, identifying micro-expressions is typically a challenging task requiring considerable effort.
- The appearance of FMiE is exhibited due to the activation of facial muscles in response appropriately to an emotional stimulus (Ekman & Friesen, 1978).
- The involved brain sections include the amygdala, insula, and prefrontal cortex which form the limbic system. These sections are associated with perception and processing emotions that trigger the final expression (Phan et al., 2002).

In conclusion, Micro FER is a growing research area that deals with identifying true human emotions and intentions through the changes in facial muscle movement (Adegun & Vadapalli, 2020). Some scholars and researchers focus on exploring the subtle and fine facial expressions that last for a fraction of a second which are known as micro expressions due to their significant role in reflecting the genuine authentic internal emotional states.

2.2.2 Macro expressions

Facial macro-expressions (FMaE) are characterised as voluntary, long-lasting, prominent, and clearly visible and can be easily discerned without the need for professional training.

The seven universal expressions of happiness, anger, contempt, disgust, surprise, fear, and sadness are typically considered to be conveyed by FMaE. Voluntary, consciously formed FMaE are generated by the primary motor cortex through the pyramidal tract (Rinn, 1984). The distinguishing factors between FMaE and FMiE can be summarised by the duration of their display, the specific combination of AUs involved in forming each expression, and their level of intensity (Zhao et al., 2022). The last-mentioned characteristic, the intensity, is low in micro expression compared with macro expression (Xia et al., 2020), and that is due to the rapid muscle movements which results in macro expressions being more obvious and recognized by the naked eye.

Universal expressions are considered a communication tool between various communities and societies. There are seven types of universal expressions:

1. Happiness: This expression is exhibited with a combination of (Check Raiser (AU6), Lid Tightener (AU7), Lip Corner Puller (AU12), Lips Part (AU25), Jaw Drop (AU26)).
2. Sadness: Revealing this emotion by a group of Inner Brow Raiser (AU1), Brow Lowerer (AU4), Check Raiser (AU6), Lip Corner Depressor (AU15), Chin Raiser (AU17).
3. Anger: The furrowed emotion expresses through these Action Units of Brow Lowerer (AU4), Upper Lid Raiser (AU5), Chin Raiser (AU17), Lip Lightener (AU23), Mouth Stretch (AU27).
4. Surprise: The signs of this feeling appeared through (Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Upper Lid Raiser (AU5), Lips Part (AU25), Jaw Drop (AU26)).
5. Fear: This bad uncomfortable feeling is shown by a combination of facial expression of Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Brow Lowerer (AU4), Upper Lid Raiser (AU5), Lid Tightener (AU7), Lip Stretcher (AU20), Lips Part (AU25).
6. Disgust: This strong emotion that shows dislike situation through Lid Tightener (AU7), Nose Wrinkle (AU9), Lips Part (AU25), Jaw Drop (AU26).
7. Contempt: This feeling is expressed by Lip Corner Puller (AU12), Dimpler (AU14).



Figure 2.3 Intensity motion difference between micro and macro expression (Allaert et al., 2019).

As can be seen in Figure 2.3, the facial segments involved in FMaE have a wide range on the face, while FMiE covers a smaller area. Most facial attributes appear in high intensity and form FMaE due to the powerful contraction in facial muscles associated with the smile line, vertical lip lines, marionette lines, nasolabial fold, chin wrinkles, eye wrinkles, and forehead wrinkles. On the other hand, the regions associated with FMiE contribute to subtle, unobvious muscle movements (Allaert et al., 2019).

Manual statistical methods are more time-consuming and less accurate in measuring FMiE than automatic systems using machine learning (Zhang, 2024). An evaluation tool to measure the intensity of facial expressions known as the micro expression training tool (METT) has been developed by Ekman for statistical measurements. The availability of a source of data in a database generated to mimic realistic data facilitates the development of automatic methods that achieve higher performance than manual approaches.

Recent research and development of DL methods is reflected in the precision of outcomes as high as 95%, while the accuracy of results from statistical methods obtained manually by a psychologist does not exceed 50-60%. Taking advantage of the significant achievements in FER that successfully identify fleeting FMiE, this thesis aims to design a method with superior performance than manual methods.

2.2.3 The Design of an Automatic Micro-FER

This thesis contributes to the indication of deterioration or recovery of patients based on the appearance of distinct combinations of FMiE, building an automatic model that can detect and identify any specific texture of involuntary micro-expressions.

This study presents two models using two kinds of DL algorithms. The proposed system monitors virtual avatars in simulated clinical wards by recognizing facial patterns from live-stream recordings, sending either normal, positive, or critical alert signals to assist healthcare professionals in their decision-making process.

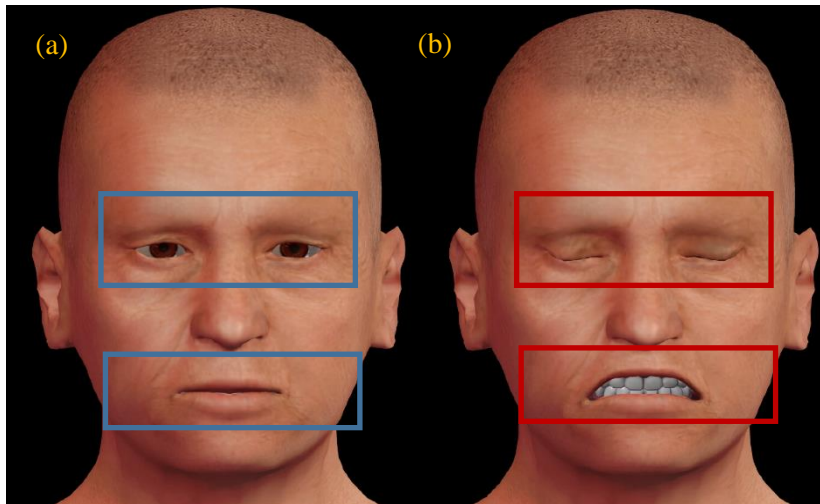


Figure 2.4 Facial expression areas that reveal if the patient is under deterioration or not.

(a) The left avatar expresses a neutral expression.

(b) The right avatar reveals deterioration status in the final stage.

The occurrence of FMaE and FMiE mostly appears on two facial segments, namely the forehead and lower face, as shown in Figure 2.4. The forehead includes the brows and eyes regions, while the lower face includes the mouth, lips, and jaw.

The designed models must be able to recognise the beginning and ending of facial action units (AUs). Five classes of AUs are markers of a patient's health deterioration, and their absence characterises recovery. In the normal stage, which is known as the offset stage when the patient is in a stable medical condition, there is no motion intensity in the five classes of AUs or FMiEs.

In the onset stage of deterioration, the facial muscles start relaxing, with minimum intensity of subtle changes in the texture of facial patterns. In the middle stage, these patterns become more visible with a higher intensity. In the final stage of deterioration, FEs reach the peak of visibility and intensity with fully expressive frame due to the maximum relaxation of the facial muscles. Some AUs of interest appear with the relaxing of facial muscles, while others occur with their contraction.

The designed AFER starts with an initial warning, indicating subtle changes in facial motion in the five classes of AU based on the appearance of features, the percentage of tension or relaxation of muscles corresponding to the five expressions, and their intensity. Any changes in facial activity will trigger the system to alert one of the three states: normal, positive, or critical.

Any changes in facial activity will trigger the system as one of the following three states:

1. Stable or normal: this means there is an absence of the existence of facial patterns of five classes.
2. Deterioration: this means the patient's medical condition becomes worse which requires immediate interference and actions from medical healthcare professionals to survive or improve the patient's condition especially if the system shows the highest intensity level of the prominence of five facial expressions.

To avoid false alarms, the appearance of AUs cannot alert the system without their visibility in both face regions and the full combination the AU set related to a particular class. FMiE usually appears along with FMaE for all universal expressions and other facial activities.

Sometimes, the changes that occur in one facial segment may obscure those in other parts of the face. However, most expressions occur in specific regions of the face. For instance, the happiness expression occurs by cheek raiser, lip corner puller, and wrinkles around the eyes due to the contraction of muscles around them (Ekman & Friesen, 1978). The inner brow raiser, lip corner depressor, and loosening of the eyelid muscles express sadness. Raising the upper lid, inner and outer brows, and opening the mouth express anger (Ekman & Friesen, 1978). Lowered eyebrows and tightened lips also represent an angry expression. Stretching and tightening the lips, along with raising the inner and outer brows, are signs of fear. A wrinkled nose pulls up the upper lip, showing the lips apart; loosened eyebrow patterns also represent fear (Ekman & Friesen, 1978).

It is important to note that any FMiE usually accompanies a FMaE (Ekman, 2009). Due to the challenging characteristics of micro-expressions, macro expression recognition still records the highest accuracy results.

2.3 Facial Expression Recognition (FER)

The full spectrum of emotions is described by various patterns of facial muscle tension. In the past few decades, various studies have contributed to analysing and understanding human emotions and their effect on human behaviours and intentions (Greche et al., 2020). FER is always considered a difficult, tedious, and laborious method due to the effort required to identify subtle changes of FE without missing them. In addition, the emotions are so different from each other due to variety in the pattern of facial muscle tension (Rinn, 1984).

Micro-expressions were first described in 1966 by Haggard and Isaac (Haggard & Isaacs, 1966) who discovered "micromomentary" expressions while scanning motion picture films searching for non-verbal communication between therapist and patient. Then, in 1969, the concept of micro-expressions was defined, clarified and expanded by Ekman and Friesen (Ekman & Friesen, 1969) who recognized them while watching a movie showing a scene of a patient with depression who showed a fleeting, pained expression.

One of the earliest works on FEs was presented by Ekman and his colleagues (Ekman & Friesen, 1971), who developed their theory by studying films of social human interactions in different cultures. Based on their analysis, they proposed the universal facial expressions for specific emotions. Seven years later, Ekman and Friesen (Ekman & Friesen, 1978) proposed universal FEs in terms of the FACS, which identified AU sets that code for specific FEs. The FACS revealed that certain AUs appear frequently in certain FEs. For example, AU6 and AU12 appear in a happy expression, while other AUs were rarely observed, such as AU9 of a disgusted expression. Furthermore, several non-predicted AUs were observed frequently in most FEs.

In 1997, Galati et al. (Galati et al., 1997) collected FEs data from participants, sighted (n=14) and blind (n=14), in response to scenarios representing six distinct emotions.

Whether the participant was sighted or blind, the FACS indicated that non-predicted AUs appeared less frequently than the theoretically predicted AUs. Later, Scherer and Ellgring (Scherer & Ellgring, 2007) implemented an experiment in which professional actors (n=12) were asked to present FEs based on scenarios corresponding a range of emotions. Here, however, FACS did not prove the existence of the large number of theoretically predicted AUs for basic and non-basic emotions.

In 2013, a study using FACS (Gross et al., 2013) stated that health professionals commonly identify expressions of sadness and fear in patients at risk of deterioration. A collaboration between North Middlesex University Hospital, University College London Hospital and the GMPR Research Group at Sheffield Hallam University proved for the first time that patterns of facial AUs can be used as indicators for admission to critical care unit (CCU). The study analysed AUs related to upper and lower regions of the face, and the direction of head tilting, using clinical metrics collected through the National Early Warning Score (NEWS) (Madrigal-Garcia et al., 2018).

Manual statistical methods for recognizing and describing FEs are heavily dependent on the subjective impressions of an observer (Rinn, 1984), so there is a necessity for automatic FER to achieve objective, accurate recognition of FEs in real time.

Although human emotions can be identified in biophysical analyses such as electroencephalograms (EEG) (Iyer et al., 2023), electrocardiogram (ECG) (Chen et al., 2021), surface electromyography (SEMG) (Chen et al., 2015), and speech signal (Wang et al., 2015), facial imaging and video analysis have been most commonly used by researchers (Sharma et al., 2023). Consequently, as the world witnesses a technological revolution and significant developments in artificial neural networks (ANN), AFER has become a hot topic in the human-computer interaction (HCI) field. Achieving highly accurate prediction by employing advanced technologies such as DL, empathetic machines can detect conscious and unconscious human feelings (Sharma et al., 2023).

The work prior to this thesis investigated various advanced technologies as a context for the design of automated models that apply ML and DL methods to recognize and classify FEs for the assessment of a patient's health status.

2.4 Automatic Micro Facial Expression Recognition (AFER)

The various fields of computer vision, human-computer interaction (HCI), and artificial intelligence (AI) have been applied in intensive studies to automate FER (Sharma et al., 2023).

Recent developments in ML and DL algorithms present new dimensions in human-computer interaction (HCI) applications (Cowie et al., 2001; Bartneck & Lyons, 2007; H. Yang et al., 2018) such as robotics, virtual reality (Hickson et al., 2019), gaming (Zhan et

al., 2008; Mourao & Magalhaes, 2013), augmented reality (Chen et al., 2015), education (Ko, 2018), advanced rider assistant systems (ADASs) (Assari & Rahmati, 2011), and digital marketing, improving the interaction between humans and computers. FER for both FMaE and FMiE has also been revolutionised by these algorithms. (Gogić et al., 2020).

However, designing and developing automatic FER is a challenging task due to the subtlety and transience of facial muscle movement (Zhi et al., 2019). Researchers have presented deep learning models that are trained properly with the available facial datasets to discern facial expressions. These trained models are then applied to unseen or new datasets under identical circumstances to validate performance.

However, DL-based FER techniques struggle with challenges, such as insufficient datasets and the subtle nature of facial motion. Consequently, various datasets, including images and videos, have been generated over the past few decades with various FEs representing different ethnicities, genders, ages, and numbers of participants (Sharma et al., 2023).

In what follows, this thesis summarizes recent automated methods exclusively focused on automatic recognition. In recent decades, AFER methods have been intensively employed in various applications. A study proposed by (Chang & Chen, 1999) used six AUs as an input vector for two neural network (NN) classifiers known as the radial basis function network and the multilayer perceptron network. Both models successfully discerned FEs with a recognition rate around 92.1%.

In 2015, Huang et al. (Huang et al., 2015) presented a micro-expression recognition framework using spatiotemporal facial representation and employing local binary patterns to extract the facial motion features and explore fine facial expressions. One year later, Jaiswal and Valstar (Jaiswal & Valstar, 2016) presented a combination of CNN and bi-directional long short-term memory (BiLSTM) that can detect facial AUs. In 2017, Sang et al. (Sang et al., 2017) presented CNNs that can recognize FEs, in which the output layer comprised seven neurons that labelled according to seven expressions, aiming to classify each image as one of the universal facial expressions. The study of Chen et al. (Chen et al., 2017) presented a CNN that uses a convolution kernel for feature extraction and max pooling operation to minimise the dimensions of the extracted features. In this work, the automatic FER analysis was also constructed to identify each facial image as one of the seven facial expressions.

Al Tae, & Jasim (Al Tae & Jasim, 2020) have used the facial design CNN model to label each face as one of the seven universal emotion categories in the JAFFE database. CNN

was trained with different grey-scale images and the accuracy of the results was 100%. The work of (Mohan et al., 2021) introduced a deep convolution neural network (DCNN) to capture and recognise patterns of FEs. The approach involved extracting local features such as edges, curves, and lines from the human face using a gravitational force descriptor, which was fed into the DCNN model to explore holistic features. In 2023, Zheng and Blasch (Zheng & Blasch, 2023) presented facial micro-expression recognition using a hybrid CNN & LSTM model, and a transformer network to capture and recognise micro expressions.

In the medical sector, use of machine learning techniques to identify FEs has made a significant contribution to correlating facial AUs and well-being. For instance, in 2019, the study of (Xu & De Sa, 2020) explored the use of computer vision methods to automatically detect pain through facial AUs, employing transfer learning methods to achieve an accurate and robust pain detection system.

The FER approach proposed in this thesis applies AI algorithms in the detection of AU sets and FEs that describe deteriorating health. Challenges to address include variations in illumination, occlusions, and features of individuals in terms of attributes such as age, gender, ethnic background, and personality.

In this respect, FER methods have progressed in two main directions, namely feature extraction and ML classifiers, or DNN. ML algorithms rely on feature extraction methods before feeding the raw data to a classifier as traditional ML models do not have the capability to recognise non-linear complex patterns. Therefore, effective extraction of relevant features is crucial for ML-based methods.

The models in this study are constructed using three main approaches. In the first, the face is detected, and facial landmarks are identified. The extracted feature vectors are fed into a NN classifier such as support vector machines (SVMs) (Shan et al., 2009), AdaBoost (Wang et al., 2004) and hidden Markov models (Uddin et al., 2009).

The combination of feature extraction and a ML method can be powerful for FER. Exploiting a pre-trained DL model to automatically extract high-level features from facial images and then feeding these features into traditional ML classifiers such as SVM and RF can have a significant impact on model performance and accuracy.

The second approach uses deep neural networks (DNN), yielding highly accurate predictions in various computer vision tasks from large scale datasets (Ko, 2018). A key advantage of these algorithms is the potential to automatically learn hierarchical representations of features directly from raw data without the need for methods to select

features from raw data, with the condition that appropriate DL algorithm must be selected for a given task (Jeong & Ko, 2018).

A variety of versions of DL models have been employed, including CNNs, LSTM, generative adversarial networks (GANs) (Yang et al., 2018; Zhang et al., 2018), and inception and ResNet modules (Hasani & Mahoor, 2017), depending on the type of task and dataset (Jeong & Ko, 2018).

Improved results have been reported for DL-based FER methods due to their ability to construct discriminative features in the learning process (Assari & Rahmati, 2011). DL models recognise combinations of AUs that describe a specific emotion rather than determining each facial feature separately (Zhao et al., 2016; Liu, Li, Shan, & Chen, 2015). For instance, DNN can recognise AU sets like AU1, AU6, AU12, and AU14 as a happiness expression, or AU9, AU15, and AU16 as a disgust expression.

Although DL AFER systems demonstrate high performance and accuracy, they have limitations related to the availability of large diverse datasets, processing time and memory use due to the multitudinous parameters in the training stage to update model weights. Decoding FEs is based on recognising and interpreting set of AUs correctly. However, as many FEs involve micro expressions whose AUs are difficult to identify precisely, decoding can depend only on the appearance features of the expression (Liu et al., 2015; Zhao et al., 2016). However, incorporating additional feature extraction methods or pre-processing steps may still be beneficial, depending on the nature of the problem and the characteristics of the data. Most recent works on FER have concentrated on using automatic recognition using images and video frames in frontal views of faces. Some of these papers concentrate on identifying spontaneous FEs in response to particular stimuli. Facial expressions can be posed or spontaneous, however, and this thesis uses the posed method as the dataset is generated through avatars.

The methodology of design, implementation, and development of an AFER using posed facial expression consists of the four primary stages of pre-processing, face detection, feature extraction, and facial expression classification. Each of these stages is systematically reviewed in the following sections. Any software project needs a development model that meets all the requirements for each stage of the project.

This thesis uses the iterative waterfall model as a software development model as described in Chapter 1. The following iterative waterfall model shows the main project stages:

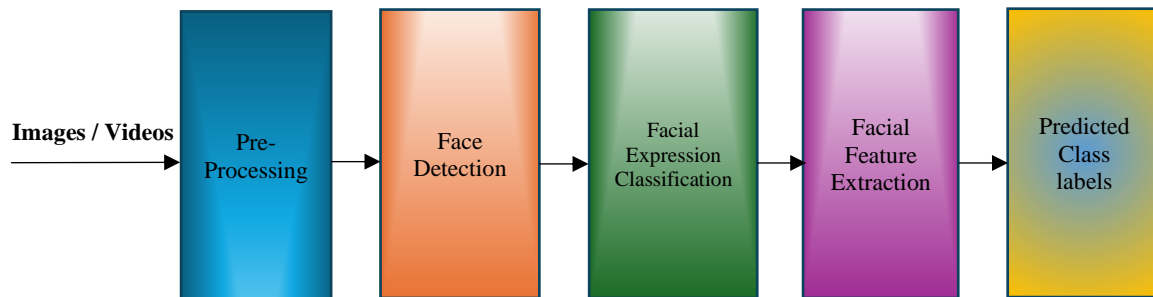


Figure 2.5 Basic framework of automatic facial expression recognition.

This study uses the iterative waterfall model as a software development model in all stages of the project. Although DL-based FER methods yield highly accurate results, limitations related to memory use and processing time can adversely affect performance in real-time tasks (Jeong & Ko, 2018).

2.4.1 Pre-Processing

While the use of cameras to record facial videos and micro-expressions is affordable, investigating each frame of a video is tedious and laborious. For acceptable accuracy, a qualified professional must inspect a video, frame by frame, to avoid missing crucial information. Many factors affect image quality, such as the pose of the face or illumination, and can have an adverse impact on the degree of accuracy. These factors can be controlled by pre-processing methods to mitigate their effect on the raw data.

The first stage of system design is data preparation, which includes pre-processing of data to achieve a quality suitable for the recognition process.

Data can be optimised using the following methods.

1. Effective lighting. When light is insufficient, this can be achieved with infrared cameras.
2. Face detection and tracking. A frontal view head pose can be detected and, if the face is not in frame; the camera will automatically adjust to the appropriate position.
3. Suppressing redundant and undesirable attributes of the dataset.

Hence, data improvement can be achieved employing pre-processing techniques such as face detection, face alignment and segmentation, frame normalization, motion magnification and data augmentation. This study explores patients' health trends by detecting it as stable or deteriorating, based on facial expressions. By enhancing the quality of input data, pre-processing improves the accuracy of relevant feature extraction for classification and leads to more accurate and robust AFER.

The following are common pre-processing steps for AFER.

- **Face Detection and Alignment.** Detecting and aligning faces in images or videos ensures excluding unwanted data and background noise as only facial regions are considered for analysis. This step helps remove background clutter and standardizes the position and orientation of faces across different images.
- **Image Resizing and Normalization.** Resizing images to a standard size and normalizing pixel values can help ensure consistency and facilitate model training. Normalization techniques such as mean subtraction and standardization can help improve the convergence of deep learning models.
- **Grey-Scale Conversion.** Converting colour images to grey scale can simplify processing and reduce computational overhead while preserving essential facial features for expression recognition.
- **Noise Reduction.** Applying filters or denoising techniques can help remove noise and artifacts from images, improving the quality of input data.
- **Histogram Equalization.** Histogram equalization techniques can enhance the contrast and visibility of facial features, making it easier for algorithms to detect and extract relevant information.
- **Data Augmentation.** Augmenting the dataset by applying transformations such as rotation, scaling, and translation can help increase the diversity of training data and improve the generalization ability of the model.
- **Pretrained Models.** Using pretrained DL models for face detection and feature extraction, such as OpenCV's Haar cascades, or DL-based face detectors like Multitask Cascade Neural Network (MTCNN) or Dlib, can streamline the pre-processing pipeline and improve performance.

Face detection is the first essential stage in FER. Paul Viola and Michael Jones (Viola & Jones, 2001a) described a face detection method known as a Haar Classifier. This technique can identify objects and attributes of faces in images and videos. The method required a large number of positive and negative face images. Positive images show faces with clear features while negative images show any objects except faces to train the model. Haar Classifiers depend on the four features of edge, line, centre, and diagonal (Viola & Jones, 2001b). The line feature recognises features that appear as lines such as nose bridge, while edge features describe edge objects such as eyebrows which usually appear darker. Hence, this method can crop out the face, ignoring the rest of the image (Viola & Jones, 2001b). In FER, employing aligning algorithms can eliminate different attributes between images and frames such as scaling and posing, improving system performance by filtering out unwanted attributes. In 2022, Febrian et al. (Febrian et al., 2022) proposed a hyper-bidirectional LSTM-CNN model and conducted research on the Extended Cohn-Kanade (CK+) database. They mitigated the overfitting issue and improved the models' performance after employing data augmentation techniques. Their findings indicate that the BiLSTM-CNN model achieves a state-of-the-art accuracy rate compared with previous models.

2.4.2 Feature Extraction

The aim of this stage is to extract stable and optimal features (Sharma et al., 2023). Applying optimal relevant attributes of facial images as references for the five classes of facial expression has a significant impact on model performance, including the speed and accuracy of the classifier.

An effective feature extraction technique used for filtering and extracting is converting data into reduced form, reducing noisy and undesirable data that represent external conditions such as lighting conditions and motion blur (Sharma, 2022).

Selection of one or multiple features is influenced by several factors, such as task requirement, the nature of the dataset, and the dimensionality between the input dataset and the targets. Selecting desirable and optimal features helps to reduce inter-class variation and interference between different classes (Sharma, 2022).

As the feature extraction techniques play a crucial role in FER systems, many computer vision techniques have been adapted by researchers to enhance performance and accuracy

of their proposed models. Feature extraction is a fundamental step in many machine learning and signal processing tasks, including image processing, natural language processing, and audio analysis. It involves transforming raw input data into a set of representative features that capture relevant information for the given task (Guyon & De, 2003). Overall, feature extraction plays a crucial role in pre-processing raw data and transforming it into a suitable representation for subsequent analysis and learning tasks. Effective feature extraction methods can significantly impact the performance and interpretability of machine learning models (Hastie et al., 2009). Feature extraction methods employed in FER include facial recognition, facial analysis, and face detection (Zeng et al., 2007a).

Effective recognition of discriminative relative features has a significant impact on a model's accuracy and performance. Features can be extracted either manually or by exploiting a deep learning network method (Jeong & Ko, 2018). There are various manual methods based on feature representation by appearance and geometric features. Appearance features describe the pattern of FEs using various feature descriptors, like a histogram of oriented gradients (HoG) (Orrite et al., 2009; Ouyang & Sang, 2013; Greche & Es-Sbai, 2016), local binary pattern (LBP) (Shan et al., 2009; Carcagnì et al., 2015; Zavaschi et al., 2013; Luo et al., 2013), scale invariant feature transform (SIFT) (Carcagnì et al., 2015; Barroso et al., 2013), and Gabor filter-based texture information (Yang et al., 2018; Zavaschi et al., 2013).

On the other hand, geometric features represent the facial landmark positions and the relationships between them (Jeong & Ko, 2018). Geometric feature methods employed in computer vision and pattern recognition focus on capturing shape-related data, like arrangement, orientation, size, and spatial relationships of features. Nowadays, different feature extraction methods have been used and combined to successfully capture essential discriminant features due to subtle changes in FEs (Sharma, 2022).

The work of Zavaschi et al. (Zavaschi et al., 2013) presents a novel method for AFER using a Gabor filter and local binary pattern (LBP) as feature extraction methods, then using SVM as a ML classifier, and genetic algorithms (GA) to identify the optimal ensemble and to reduce the error rate and size of the ensemble. In the same year, the work of (Luo et al., 2013) introduced LBP for feature extraction from greyscale images of mouth parts, and SVM to categorise FEs.

Three years later, (Greche & Es-Sbai, 2016) described an AFER using HoG as a feature extraction method and normalized cross correlation, reporting an accuracy of around 83.6% for five FE in images of varying resolution and colour space. Carcagni et al. presented a comprehensive study on the role of HoG descriptor in FER and described the appropriate set of HoG parameters that can improve this descriptor to classify FEs.

Perikos et al. (Perikos et al., 2018) described an AFER system conducted on the Japanese Female Facial Expressions (JAFFE) database to recognize FEs using adaptive neuro fuzzy inference systems. In this method, faces are extracted using the Viola-Jones algorithm, then FEs are analysed and the deformations in face parts like eyes, eyebrows, and mouth are identified. Then, attributes including position, length, width, and shape are captured. Representing the FE deformation, these feature vectors are fed into adaptive neuro fuzzy inference systems to capture FEs. This model showed encouraging results, registering 90% average accuracy in FER.

The study of (Choi & Oh, 2006) proposed a real-time FER technique with a statistical model known as the active appearance model (AAM), usually used in computer vision and image processing, with a second order minimisation and an NN. Second order minimisation improves correct convergence with AAM with a minimum loss of frame rate. Hence, the facial shape is extracted correctly with AAM, reducing errors in FER. The high dimensional feature vectors of six FEs, involving relative patterns and facial components, were classified by a multi-layer perceptron, achieving an excellent recognition rate of over 99%.

The paper of (Tanchotsrinon et al., 2011) describes AFER in two main stages. First, points in the face region were located to form graph-based features and, second, the NNs were trained to recognize the FE from the corresponding feature vector. In the first phase, 14 points were manually located to create a graph with edges connecting the points. The method achieved 95.24% accuracy.

In 2011, Tsai et al. (Tsai et al., 2011) introduced a new distance-based feature extraction method that works over different domain datasets of various classes, samples, and dimensions, and pattern classification tasks. This method depends on extracting two types of distances, either that between the data and its intra-cluster centre, or that between the data and its extra-cluster centre using naïve Bayes, k-NN, and SVM classifiers. The distance-based features could enhance classification performance and accuracy, especially for datasets of a smaller numbers of classes, number of samples, and dimensionality of features.

The work of (Suk & Prabhakaran, 2014) proposed an AFER based on machine learning, specifically a set of SVMs, to recognize six FE along with a neutral expression. The FE features were extracted by an active shape model (ASM) placing landmarks on the face. Then, new dynamic features were generated by replacing neutral features with other FE features. The classifier model shows around 86% accuracy using a 10-folds cross-validation method to classify 309 samples using Extended Cohn-Kanade (CK+). This model was applied on a Samsung Galaxy S3 with 2,4 fps, and the accuracy of the real-time FER app was 72%.

Traditional feature extraction approaches are suitable for real-time tasks due to the speed of the model in learning extracted features, and these methods work effectively with small-scale datasets. However, in terms of performance, they are inferior to DL techniques (Jeong & Ko, 2018).

Mollahosseini et al. (Mollahosseini et al., 2016) presented a DL architecture to analyse and recognize FE across various face databases, like Multi-PIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. The model consists of two convolution layers, each followed by a max pooling layer, and output results are fed to four fully connected inception layers. The model records high performance with accuracies comparable with traditional CNN architectures of around 93.2% for CK+ and 77.6% for the MMI database (Pantic et al., 2005a; Valstar & Pantic, 2010). Hasani and Mahoor (Hasani & Mahoor, 2017) proposed models comprising 3D Inception-Resnet followed by an LSTM unit to capture spatial features within input images along with temporal relations between features across frame sequences. They used facial landmarks as features rather than focusing on facial regions which provide irrelevant data not contributing to FEs.

Liu et al. (Liu et al., 2015) proposed a system using facial databases CK+, MMI, and FERA. Deformable action parts were incorporated into a 3D CNN model which can recognise articular facial action parts that represent spatial features. The model achieved state-of-the-art FER accuracy.

Generative adversarial networks (GANs) are a type of DL model introduced by Goodfellow and other contributors in 2014 (Goodfellow et al., 2014), and have shown satisfactory results by combining elements known as a generator and a discriminator. Zhang et al. (Zhang et al., 2018) proposed an AFER model based on GAN using the Multi-PIE and Static Facial Expressions in the Wild (SFEW) datasets (Dhall et al., 2011) datasets. The model was trained and evaluated using various head poses and facial expressions. Additionally,

the proposed model automatically generates facial images along with different FEs in different poses to expand and enrich the training dataset. Evaluation of the GAN model showed an accuracy around 91.8% for the Multi-PIE and 26.58% for the SFEW dataset.

The AU-based method is different from other approaches using overall face features, relying on pre-defined AUs to code for specific expressions based on FACS. In recent years, AU-based methods have been used in the DL approach. Zhao et al. (Zhao et al., 2016) adapted deep region and multi-label learning (DRML) method to identify AUs and capture FEs based on face alignment. This model achieved highly accurate AU detection including the correlations between AUs. However, the results depended on face alignment and equal treatment of blocks, which may result in eliminating certain regions that could have improved results. This model recorded the highest average F1-score and AUC for BP4D and DISSA datasets compared with other approaches.

Liu et al. (Liu et al., 2015) presented AU-inspired deep networks (AUDN) to identify the AU sets that form facial muscle movement and FEs. This model involved three main processes. First, a micro action-pattern (MAP) was learned by the constructed model which consisted of a convolution layer followed by a max pooling layer to capture informative local appearance variation. Second, the optimal feature grouping was determined to integrate correlated MAPs. Finally, a multilayer learning process was used to produce high level representation for FER. The results showed 93.7% accuracy for the CK+ and 75.85% for the MMI database. By applying linear classifiers for learned features, the model records excellent results on all the databases, validating the proposed model in lab-controlled and natural settings.

While DNN-based FER is one of the latest advancements and delivers exceptional performance, this technique remains resource, memory and cost-intensive. Consequently, traditional categorizing methods are under continued study for integration into real-time tasks due to their lower computational complexity and high level of precision. As feature extraction techniques scale down raw data, it is essential to keep effectively depicting their source. Feature extraction methods can be generally categorised as appearance or geometric feature extraction (Jeong & Ko, 2018).

Appearance feature methods extract texture patterns from a face, including image intensity, gradient, image filters, while geometry feature techniques exploit geometric relationships of facial attributes, including facial landmark positions, angles between certain points, and Euclidean distance. Although appearance features yield better performance than geometric

features, the latter demonstrate greater resilience and robustness to subtle alterations and variations in face position, scale, size, and face direction. Combinations of these approaches, hybrid approaches, can be employed to leverage complementary information for improving FER performance (Sharma et al., 2023).

In video stream consisting of frame sequences, feature extraction methods known as spatial-temporal approaches simultaneously examine spatial features and FEs in a frame and the temporal features within sequences (Sharma, 2022).

Feature extraction and representation approaches are an essential step for data mining and pattern classification tasks since the extracted features have a direct and significant impact on the classification accuracy (Tsai et al., 2011), as they extracted relevant information from facial images or videos for subsequent classification. In summary, deep analysis and investigation of features that influence classification could provide deeper insights into a model's behaviour and indicate potential areas for further improvement.

2.5 Upper and Lower Parts of the Face

Previous studies report FER for the entire face, like the work of Prkachin and Solomon (Prkachin & Solomon, 2008) on identifying pain intensity. However, in critical care units, patients' faces may cover with a mask due to breathing issues, obscuring their facial expressions (Yuan et al., 2022). Indeed, work presented by Roberson et al. (Roberson et al., 2012) showed that extracted features for FER are reduced when a portion of face is not observable, reducing the accuracy of emotion recognition. Gori in 2021 (Gori et al., 2021) reported similar findings.

Recent studies that have examined masks and facial emotion recognition have found that wearing a mask reduces the accuracy of emotion recognition (Mukhiddinov et al., 2023). Another work proposed by Gori in 2021 (Gori et al., 2021) show that face mask has great impact on the ability of observing facial expression and recognizing emotions (Gori et al., 2021). Other work presented by (Carbon, 2020), (Gülbetekin et al., 2023), and (Pazhoohi et al., 2021) have shown the impact of masking the face. However, some research has succeeded in identifying and recognizing certain FEs using partially covered facial images. For example, Yuan et al. (Yuan et al., 2022) designed and trained models using the UNBC-McMaster dataset to capture features that identify pain and pain intensity from masked faces based only on AUs in the upper part of the face using Swin-Transformer. The model

achieved around 90% accuracy.

Although some facial expressions such as pain can be successfully identified with or without a face mask (Carbon, 2020; Yuan et al., 2022), still measuring and recognizing facial expressions with face mask minimizes the accuracy of predicting human emotions (Mukhiddinov et al., 2023).

Some facial expressions such as happiness appear intensively through AUs in the lower part of face (mouth region), while others like fear are mainly located in the upper part of face (eye region) (Ekman & Friesen, 1976). Smith et al. (Smith et al., 2005) used the bubble technique proposed by (Gosselin & Schyns, 2001) to estimate how the brain processes and identifies which face regions are correlated with categorization of six universal expressions. Studies based on bubble technique (Gosselin & Schyns, 2001) show that the mouth segment is the most informative for universal facial expressions such as happiness, surprise and disgust, the upper part of the face including the eyes provides the most detailed features for fear and anger, and both segments are informative for sad and neutral expressions. Hiding half of the face image presenting only the upper or lower part results in similar findings. As stated above, the AUs of happiness and disgust expressions are recognisable from the lower part of the face, while features of anger, fear, and sadness are most obvious in the upper part; therefore, these segments are informative for sad and neutral expressions (Smith et al., 2005; Blais et al., 2012; Wegrzyn et al., 2017).

Hiding half of the face image presenting only the upper part or the lower part results in similar findings. For example, as stated above the action units of happiness and disgust expressions being most recognizable from the lower part of the face, while features of anger, fear and sadness being most obvious from the upper part of emotional face (Calder et al., 2000).

In 2012, Roberson et al. (Roberson et al., 2012) addressed the impact of sunglasses and masks on FER. The face mask condition was a non-realistic grey ellipse added to the mouth area, not covering the nose and cheek regions of the presented facial images. When sunglasses were added to the facial images, their results indicated a reduction of accuracy in FER for happy, sad, surprise, fear and anger emotions. FER accuracy was further reduced when the grey ellipse was added. However, FER accuracy for the masked face alone was not reported. In another work by Noyes et al. (Noyes et al., 2021), FER for faces occluded with a face mask had a higher error rate than for those with sunglasses. In 2001, Gosselin and Schyns (Gosselin & Schyns, 2001) introduced a new technique known as bubbles,

describing various impacts on FER for all FEs when occluding the lower part of face, including mouth, cheeks, and nose regions, with face masks. Furthermore, studies such as (Beaudry et al., 2014; Schurgin et al., 2014) showed that the outcomes of classifying facial expressions are various regarding covering the eye against the mouth regions.

Overall, it has been demonstrated that covering the upper part of the face with sunglasses has a negative impact on FER for happiness, while occluding the lower part of the face with a face mask disrupts FER for happiness, disgust and anger more than hiding the upper part of face. At the same time, other emotions gave varying results. For instance, (Kotsia et al., 2008) observed that covering the mouth interrupted FER for disgust and anger more than covering the eyes, while the work of (Schurgin et al., 2014) showed the opposite.

2.6 Face Detection using MediaPipe

The initial stage of pre-processing is the detection and tracking of the face. This is an important research field but beyond the scope of this thesis, which applies pre-existing methods and algorithms.

From 2012, CNNs have provided many breakthroughs in the computer vision field solving a variety of problems from image analysis to image classification (Krizhevsky et al., 2012a). In various CNN architectures, face detection is affected with a feature extractor like S3FD (Zhang et al., 2017), PyramidBox (Tang et al., 2018), DFSD (Li et al., 2019), which have achieved state-of-the-art face detection (Sutanto et al., 2021). The transfer learning from a pre-trained CNN model includes two techniques: feature extraction and fine-tuning, that are ordinarily utilized for best outcomes (Darwish et al., 2020).

Although these CNN-based face detection methods are robust to the large variation of facial appearance, they are too time-consuming for real-time performance, especially on CPU devices. Research on face detection is focusing on decreasing model size to run in real-time on low computational power devices such as CPU (Sutanto et al., 2021). Multitask Cascade Neural Network (MTCNN) (Zhang et al., 2016) and FaceBoxes (Yang et al., 2015) are examples of real-time state-of-the-art face detection methods (Sutanto et al., 2021).

Transfer learning from a pre-trained CNN model includes the two techniques of feature extraction and fine-tuning that are ordinarily utilized for best outcomes. Face detection methods based on ML algorithms have developed significantly achieving highly accurate results compared with other statistical methods (Al-Nuimi & Mohammed, 2021). Research at

Megvii Technology (Sitepu et al., 2021) proposed a DL-based face detection and alignment algorithm called RetinaFace that uses a single fully convolutional network architecture to simultaneously detect facial bounding boxes, facial landmarks (such as eyes, nose, and mouth), and face orientations (yaw, pitch, and roll) with highly efficient performance and accurate prediction for various poses, scales, and occlusions. Therefore, it is considered suitable for real-time apps.

A region-based convolution neural network (R-CNN) proposed by (Girshick et al., 2016) is a DL-based object detection algorithm that can be used for face detection and alignment. It is particularly well suited for face detection in complex and dynamic scenes. Dlib is a library used for computer vision and image processing applications. This method uses various tools for face detection and recognition, including facial landmark detection and face alignment (Reza et al., 2021).

A recent article by (Sheremet et al., 2023) used DLib, MediaPipe, Key-Point R-CNN, retinaFace, and HRNet for face detection, replacing faces in a video stream for a comparison among these techniques to improve the efficiency and accuracy of model.

OpenCV is an open-source library used for image processing and video analysis that is widely used in computer vision (Duan & Luo, 2022; Tirupal et al., 2023). OpenCV was introduced by Intel in 2001 and programmers have contributed to its development. OpenCV has many computer vision functions including face detection, facial recognition including facial landmarks, and face tracking (Reza et al., 2021).

MediaPipe (reimplementation of BlazeFace) (Duan & Luo, 2022; Latreche et al., 2023) is an open-source framework developed by Google that is optimized for real-time video processing. It provides high accuracy face detection and uses machine learning to improve accuracy. For facial landmarks, it can detect 468 3D key points of Face Mesh (Sheremet et al., 2023). The MediaPipe face detector operates on images or video streams. It can be used to locate faces and facial features within a frame. This task uses an ML model that works with single images or continuous sequences of frames. The task outputs face locations with the following key facial points: left eye, right eye, nose tip, mouth, left eye tragion, and right eye tragion.

Overall, MediaPipe is a powerful framework for designing real-time multimedia applications, offering a wide range of pre-trained models and tools for developers to create custom pipelines tailored to various tasks. It has been widely used due to its simplicity, real-time efficient performance, highly accurate prediction, and versatility across different

platforms and applications. For optimal performance, tracking and detection should be run in parallel, so the tracker is not blocked by the detector and can process every frame (Lugaresi et al., 2019). This thesis proposes auto-tracking face detection while ignoring other parts of the image to track patients' faces in real-time.

2.7 Facial Landmarks

Facial landmarks are the key reference locations of facial features in computer vision and image processing disciplines because they work as a standardized and referenceable approach for locating, representing, and interpreting facial features across various facial images (Wu & Ji, 2019). In recent years, many applications such as FER (Al-Tekreeti et al., 2024), face recognition (Zhao et al., 2003), face alignment (Lee et al., 2023), face tracking (Kalal et al., 2010), facial analysis (Al-Tekreeti et al., 2024), 3D face construction (Wood et al., 2022), and face detection (Al-Tekreeti et al., 2024; Chandran et al., 2024) have used facial landmarks as a critical step to define and locate features (Zafeiriou et al., 2015).

Typically, facial landmarks involve the AUs that correspond to facial muscles (Al-Tekreeti et al., 2024). Some of these landmarks are as follows and are depicted in Figure 2.6.

1. Eyes. Landmark points are often located at the corners and centres of the eyes to capture sight position, shape, size, and orientation of the eyes.
2. Nose. Landmark points are usually placed at the top and base of the nose, aligning its sides to detect its shape and orientation.
3. Mouth. Landmark identifier points are located at the corners and centre of the mouth, helping to identify its shape and expression.
4. Jawline. landmark points are marked along the jawline to detect and identify overall face shape.

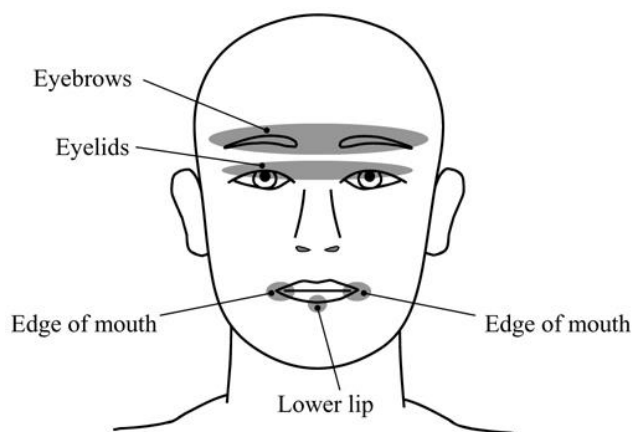


Figure 2.6 Some of main regions of interest of facial landmarks (Hachisuka et al., 2010).

When deep algorithms detect and identify these landmarks and define the spatial relationship between them, they can recognize FEs, identify a person, and predict their age and future facial features. Different methods have been employed for FER, ranging from traditional image processing techniques to higher advanced ML algorithms. Various ML algorithms have been trained over many databases of digital images, but the CNN model has consistently and robustly recognized facial features and landmarks (Krizhevsky et al., 2012b)

Many previous studies have presented various frameworks using face detection techniques. A facial landmark detection algorithm based on the cascaded regression method is described in (Kazemi & Josephine, 2014). This algorithm is implemented in the Dlib library (Savin et al., 2021). In 2016, Yan et al. (Yan et al., 2016) presented an AFER framework based on facial landmarks by using a CNN-RNN model. Facial features were extracted from sequences of frames by feeding each facial image into the finetuned VGG-Face model, and then the features in facial images were sequentially traversed in a bidirectional RNN to recognise and capture dynamic subtle changes in facial patterns.

In 2019, Kartynnik et al. (Kartynnik et al., 2019) introduced a real-time high quality predicting method for detecting 468 facial landmarks from 3D mesh representing of a human face based on NNs using a single camera. In 2021, Al-Nuimi and Mohammed measured (Al-Nuimi & Mohammed, 2021) head pose angle using trigonometric functions and facial landmarks based on a MediaPipe approach. In the same year, Singh et al. (Singh et al., 2021) reported a real-time framework that identifies human action under various conditions and viewing angles through frames. They used MediaPipe as a detection model

to provide pose, face, and hand landmarks in frames provided in real-time using OpenCV. The proposed model provided a total of 501 landmarks which were exported as coordinates to a CSV file to classify and identify body language poses.

Recent research by Sharma et al. (Sharma et al., 2023) presented an AFER for variously oriented faces based on calculating distance-based features by applying a face mesh to find inter-spaces between facial landmarks. The images were pre-processed with cropping and resizing techniques, then the facial features were normalised to explore the optimal attributes for classifying FEs. They used the IIITM Face dataset to classify FEs using ML algorithms such as SVM classifier, achieving around 61% accuracy. For the KDEP database, accuracy reached 80%.

MediaPipe is also a set of libraries, pre-trained models, and methods for different kinds of tasks, such as face identification in the image, facial landmark detection, body pose tracking, and object recognition. If a task requires ML, MediaPipe applies TensorFlow. The solution of a problem by MediaPipe leads to the construction of a pipeline for media information (video flow) processing. Pipelines exist to solve widespread video processing tasks like facial landmark detection. MediaPipe enables detection of 468 facial landmarks arranged in fixed quads and represented by their coordinates (x, y, z). The mesh topology comprised by these landmarks is presented in Figure 2.7(Savin et al., 2021). MediaPipe presents three distinct models to localize facial landmarks. The first identifies a face with facial landmarks. The second, known as the face mesh model, applies a complete mapping to the face consisting of 468 3D face landmarks. The first and second models have been applied to facial images with the results shown in Figure 2.7.

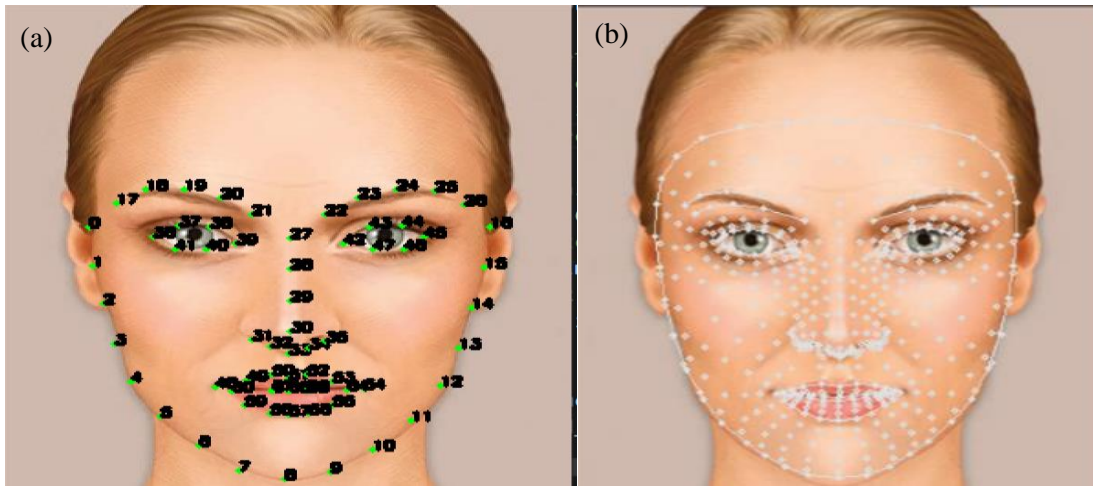


Figure 2.7 Two different facial landmarks models. (a) detect the presence of faces with key facial landmarks. (b) face mesh model applies a complete mapping of the face that consists of 468 3D face landmarks.

A third model called Blendshape uses facial landmarks to recognise facial features and FEs by working with the output of the face mesh model to predict 52 scores that are the coefficients representing various FEs. Consequently, this model is a package of the three models, including face detection as a first stage, applying full mapping of facial landmarks (468 landmarks in 3D), and, lastly, capturing and recognizing FEs through the Blendshape model.

Using facial landmarks either from a single model or from a bundle provides valuable information for exploring FER. The MediaPipe face mesh solution estimates 468 3D facial landmark points in real-time (Nazarkevych et al., 2023). This type of face mesh solution enables developers to create multi-modal cross-platform applied ML pipelines. It also provides a solution for various computer vision applications such as face detection, iris detection, hand detection, pose, holistic, hair segmentation, instant object tracking, and object detection (Khanum & Pramod, 2022). It can also be implemented on mobile devices due to its light-weight architecture. It delivers real-time performance critical for live experiences. The face mesh algorithm achieves such performance by employing two real-time DNN models simultaneously.

2.8 Conclusion

In conclusion, various approaches have been developed and applied to different components of FER systems, generally based on the facial action coding system (FACS). The constructed FER models and their constituent blocks have been evaluated here. Various techniques and algorithms used in these blocks for improving performance have also been presented. This extensive literature review describing the significant progress made in the field of AFER provides a solid comprehension of the methods to be explored in our project. The most common ML algorithms in modelling FER have been presented and applied to a generated database named Patients at Risk of Deterioration-Facial Expressions (PRD-FE) that mimics the FEs of patients under risk of deterioration. Exploring various FER has helped to uncover the current trends and challenges in this field. The study shows that both manual and DNN techniques have contributed to accomplishments in AFER analysis. Based on the review it is fair to claim that progress in DL-based FER systems was considerably influenced by the design of novel databases through various augmentation approaches. Significant studies report employing good quality images and frame sequences to achieve state-of-the-art FER accuracy.

Designing methods for FE feature representation to effectively encode subtle movements is a research area within MaFE and MiFE analysis which will be explored in this thesis. We have also conducted experiments exploring both macro and micro-expressions in FER. We propose the creation of a benchmark (PRD-FE) database, a significant contribution to knowledge that will be described in Chapter 4.

Chapter 3

Introduction to Machine Learning

3.1 Introduction to Machine Learning and Data analysis

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that focuses on designing and developing computer algorithms to facilitate and improve learning processes from data patterns. The importance of machine learning models lies in their ability to efficiently handle intricate tasks by adapting to unseen data and changes in environmental conditions. (Abo-Tabik et al., 2021). They have been widely adapted across different disciplines. (Ray, 2019). The data is like the fuel for ML models that represent the engines to power this data-driven world. Highly relevant and diverse data aid automatic algorithms decisions and predictions on unseen data over time. The core work of its models is based on the principles of acquiring knowledge through data analysis, identifying patterns, and extracting distinctive characteristics from created or gathered data (Sarker, 2021).

(Abo-Tabik et al., 2021). ML methods trained by a given set of examples refer to input data to accomplish specific tasks, and the outcomes and predictions depend on the quality and quantity of given examples. Nowadays, ML is essential for its ability to handle complex tasks and deal with enormous datasets to provide valuable insights that can drive decision-making.

ML models fall into three primary categories: supervised, unsupervised, and reinforcement learning (Muhammad & Yan, 2015). The most prevalent form of ML models is supervised learning trained using pre-defined labelled data to explore and learn the relationships between features and map them to their corresponding targets. With proper training, they acquire the potential to produce precise predictions. As an illustration, supposing a supervised algorithm is provided with raw data consisting of animal images (e.g., lion, tiger, monkey, etc.) that are assigned labels indicating their names, the model may effectively identify and categorize previously unseen animal images. As another example, in the finance and marketing field, a supervised model can predict the fall and rise in future prices from historical data. Common examples of supervised models include supporting vector machine for classification problems and linear regression for regressions tasks.

In contrast, unsupervised models are trained with an unlabelled dataset and the outcomes are unknown, so the model tries to find features and the relationships in the dataset on its own (Schmarje et al., 2021). This method is usually employed for clustering and dimensionality reduction problems (Peralta & Saeys, 2020). Clustering involves a set of similar data points together (Jain, 2010), while dimensionality is the transformation of data from a high-dimensional space to a low-dimensional space by decreasing random variables to obtain and retain a set of meaningful properties of the original data (Dash et al., 1997). Common examples of unsupervised learning methods are k-means for clustering tasks and principal component analysis for dimensionality reduction tasks (Ding & He, 2004). Unsupervised learning algorithms are applied in marketing, grouping costumers that have similar behaviours or demographic data without any pre-existing labels (Tsiptsis & Chorianopoulos, 2011).

The third type of machine learning algorithms is reinforcement learning, which is particularly suited to tasks handling sequential data. A decision is reached by interacting in real-time with the environment at each step, which affects future outcomes. Common applications of this method are games and robotics as the agents can be rewarded or penalized through maximizing or minimising their rewards based on the actions that have been taken.

Deep learning (DL) is a branch of ML concerned with computer algorithms based on multi-layers inspired by the structure of the human brain to imitate human learning, thinking, and analysis (Falavigna, 2022). Several key features distinguish the DL algorithms from other traditional ML algorithms. They are more complex, have hierarchical architecture, train and learn from huge datasets, and can predict outcomes with high accuracy (Falavigna, 2022). A wide range of applications use DL algorithms, including handwriting recognition (Zhang et al., 2020), speech-language translation (Sarmah et al., 2024), and FER (Al-Tekreeti et al., 2024). The importance of DL is its ability to handle and make accurate decisions based on vast amounts of data in real-time, driving innovation in many fields such as healthcare in predicting disease outbreaks, pain recognition, diseases recognition from symptoms, and in the finance field they have been used in fraud detection and trading algorithms. ML proceeds through seven main stages: data collection, data pre-processing, choosing and designing the model, training the model, evaluating the model, hyperparameter optimisation, prediction, and deployment.

ML models have two important components that affect the learning process and the model design known as hyperparameters and parameters (Bengio et al., 2015). The configuration of hyperparameters is set manually and optimised during the design and techniques stage of the model, so they do not learn from the dataset as they are pre-set before the start of training process. Choosing the right configuration settings for these parameters is essential for achieving optimal model performance as they influence the learning process (Bengio et al., 2015). On the other hand, the parameters are internal variables that are adjusted and optimised by learning from the data during the training stage to reduce the difference between the target results and the actual outputs. The parameters are the coefficients in the linear regression that act as the weights and biases in the NNs (Bengio et al., 2015).

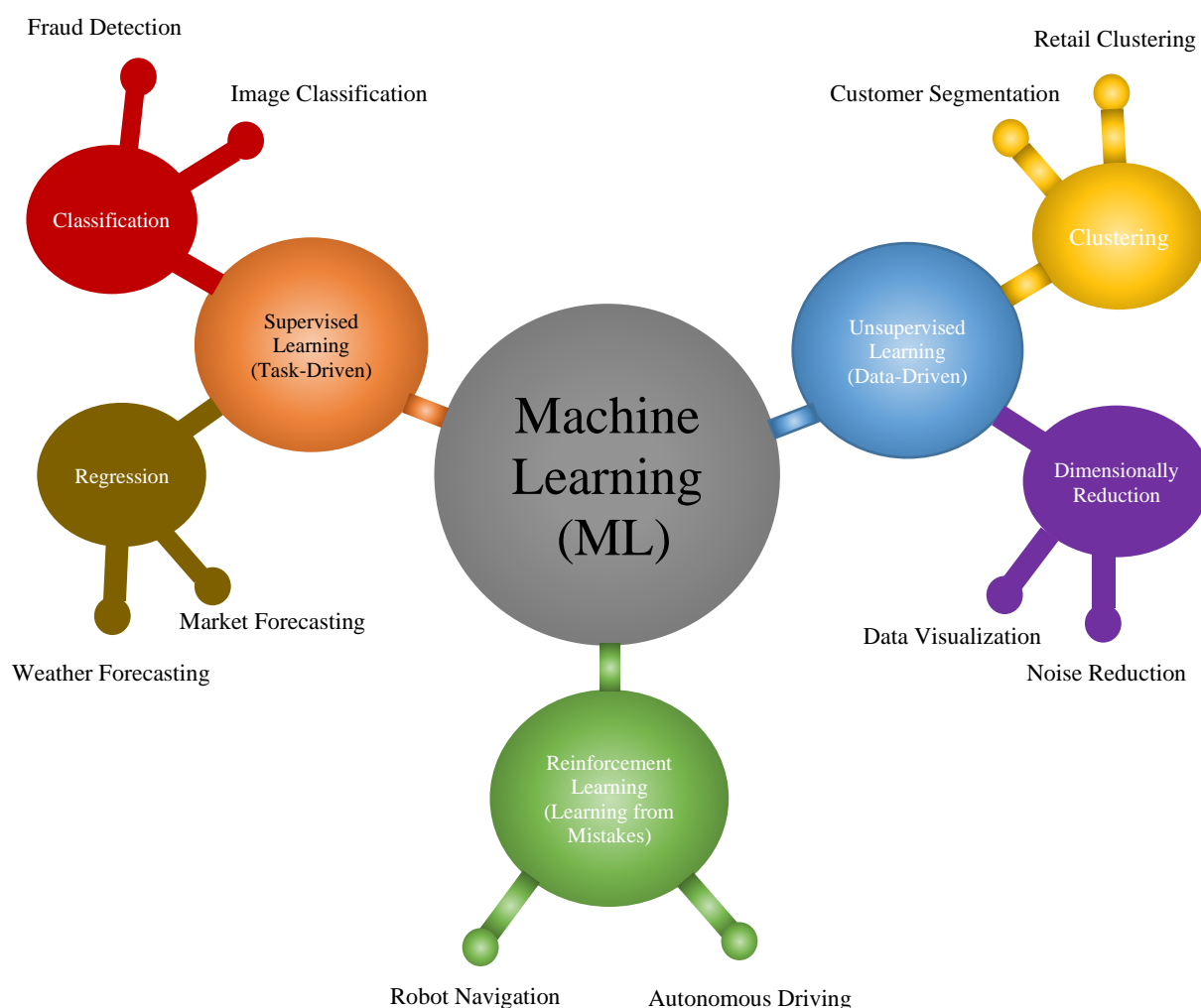


Figure 3.1 Types of Machine Learning Algorithms.

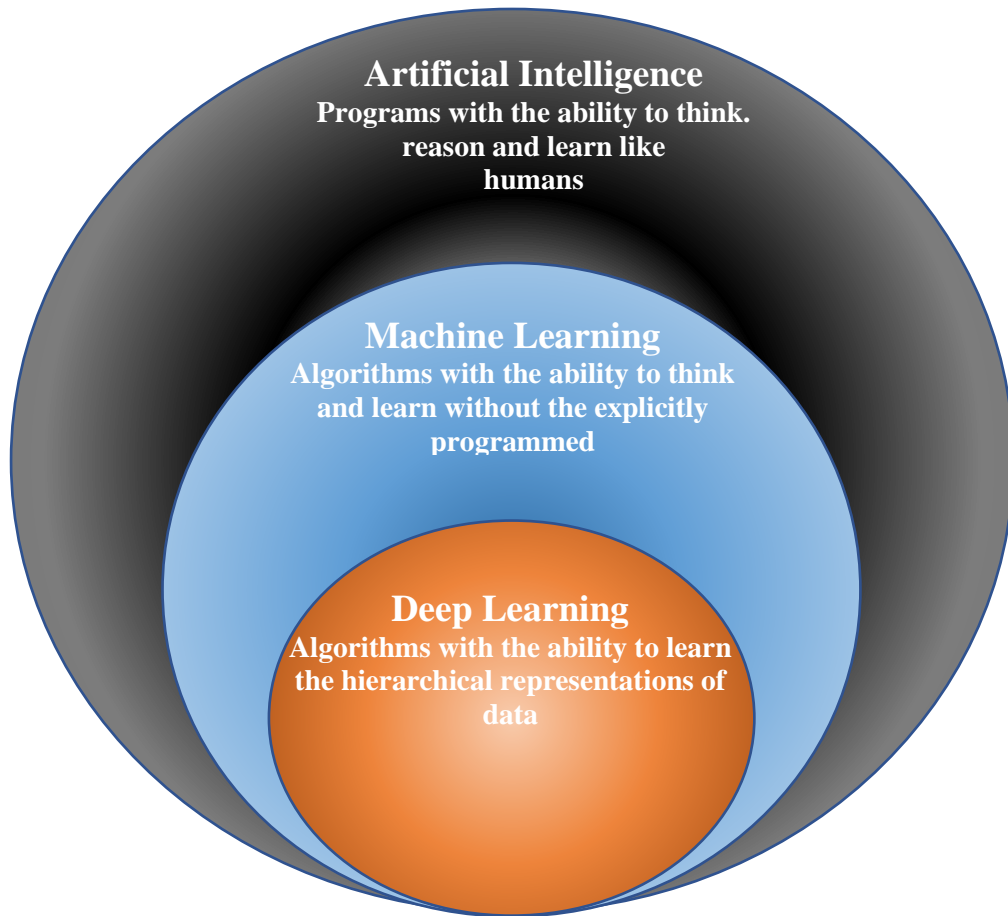


Figure 3.2 Euler Diagram showing the AI hierarchy.

After the training data stage, a model must be evaluated for its reliability by testing its ability to handle and master the target task (Bengio et al., 2015). The evaluation of ML models is based on essential metrics such as accuracy. The aim of a ML engineer or designer is to achieve the highest model accuracy, a measurement which represents the model's ability to find the features and relationships in data related to the target task. The accuracy focuses on the number of true outcomes and is calculated by comparing the number of correctly predicted samples to the overall number of predictions (Abo-Tabik et al., 2021).

There are four essential measures to evaluate the model:

1. True Positives (TP): The number of accurately predicted samples.
2. True Negatives (TN): The number of rightly predicted values as negative.
3. False Positives (FP): The number of positive samples that are incorrectly predicted.
4. False Negatives (FN): The number of false negative samples that are inaccurately predicted.

Accurate model predictions consist of true positive and true negative samples. Misleading predictions include the false negative and false positive samples. The accuracy of model can be determined by the following equation (Ikram & Cherukuri, 2016;Thaseen & Kumar, 2017):

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True negatives} + \text{False positives} + \text{False Negatives}} \quad (3.1)$$

The accuracy metric is a straightforward measurement, but it cannot be considered a sufficient evaluation for all tasks due to certain limitations. For instance, it might provide an inappropriate measurement when evaluating imbalanced classes, where there is a substantial difference in the number of samples between different classes. In this case, the metric of accuracy may be very high because it correctly predicts the majority class, even though the model performs poorly in the other minority classes.

Another metric is precision, describing the number of correctly predicted samples of positive class. It can be calculated by finding the ratio of correct sample predictions to the overall number of samples identified as positive. The proportion between true positives and the total of both true positives and false positives can be calculated in the following formula (Chakravarthi et al., 2020; Abo-Tabik et al., 2021):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3.2)$$

Recall, also referred to as sensitivity, is an evaluation metric commonly used in the medical and biological domains. It measures the ability of a model to accurately detect positive samples and is sometimes called the true positive rate. The metric is calculated by dividing the number of genuine positives by the total number of positive samples (Chakravarthi et al., 2020).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.3)$$

Another popular metric for evaluating model performance is the F1 score, which is considered the harmonic mean of precision and recall as it provides the balance between them and is an active metric in imbalanced classes. Its importance appears in evaluating a model's ability to detect true positives and false negatives. F1 Score is designed as the geometric mean of Precision or PPV and Recall or True Positive Rate (TPR) (Chakravarthi et al., 2020; Raouhi et al., 2022). The equation for calculating the F1 Score is as follows (Raouhi et al., 2022):

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.4)$$

The last common metric is specificity, the ratio between true negatives and actual negative samples. The following formula shows the way to calculate this metric:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (3.5)$$

These metrics give a comprehensive evaluation of ML model performance, which is measured by testing the model on unseen or new data during the testing process. One of the most common evaluation methods is the confusion matrix which uses the four essential components true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in assessing and evaluating the performance of classification models (Tharwat, 2018).

It is usual to evaluate the model on a subset of the dataset that has not been previously seen during the training phase. For model evaluation, the initial stage involves dividing the dataset into two parts: the training set and the testing set. It is recommended to further split the training set into training and validation sets in order to prevent any issues related to data leakage. An effective model is characterised by its ability to achieve high accuracy in both the training and testing phases, while also demonstrating robustness to variations in hyper-parameters. Nevertheless, despite the model's satisfactory performance during training, it sometimes shows a drop in its performance while testing. This scenario is referred to as over-fitting. Over-fitting typically arises when the model is excessively trained, although it can also manifest when the model is excessively intricate (Abo-Tabik et al., 2021).

In over-fitting, the model becomes over specialised to the training set and, as a result, it loses its generality and its ability to capture the relations within previously unseen data. Another factor that can make a model susceptible to over-fitting is its capture of noise accompanying the data and treating it as meaningful. Consequently, it performs poorly on unseen data due to the absence of noise data (Erickson et al., 2017).

The last step in the dataset preparation stage involves partitioning the data into two crucial sets: the training data and the testing data.

To mitigate data leakage issues, divide the training dataset into training and validation datasets. The highly qualified model should demonstrate accurate performance in identifying relevant features and recognizing relationships and patterns in training and testing datasets. In addition, the model can be considered robust when it shows high reliability for external and internal factors such as illumination and facial rotation.

Even when the model's performance shows high accuracy on the training dataset, it shows limitations and degradation in its performance through the testing process, which is known as an overfitting problem (Erickson et al., 2017; Abo-Tabik et al., 2021). Some factors, such as overtraining or high model complexity, can cause overfitting (Erickson et al., 2017; Abo-Tabik et al., 2021). Capturing accompanying noises in the data and treating them as meaningful patterns can also lead to overfitting issues in the model, hindering its performance on unseen data due to the lack of noise (Erickson et al., 2017). Already stated above.

The scarcity and size of datasets also affect the learning process because a model may memorize the instances rather than learning and recognizing the underlying features or patterns, resulting in poor adaptation to new data (Erickson et al., 2017). Hence, even if a model is a specialist in predicting during the training process, it can lose its ability to identify relevant features and meaningful patterns in unseen datasets (Erickson et al., 2017).

The cross-validation process can help in detecting and mitigating the overfitting model performance by employing statistical techniques to estimate how well the model generalizes to unseen data and to boost the accuracy of model performance by training and testing the model on various data subsets.

The cross-validation process can help detect and mitigate overfitting by employing statistical techniques to estimate how well the model generalizes to unseen data and to boost the accuracy of model performance by training and testing the model on various data subsets. The k-fold cross-validation is widely employed to solve overfitting by splitting datasets into k-folds. In this technique, the model must be trained and evaluated k times on various portions or folds (Abo-Tabik et al., 2021). In each iteration, the model is fed with a different fold of the validation dataset and trained with the remaining folds. In addition, this method helps to overcome underfitting in which the model suffers from poor performance on seen and unseen data during the training and testing phases. Furthermore, it reduces data variability because each fold of the input dataset will be exposed to the validation and testing process (Xiong et al., 2020). However, models that use this method are time-consuming, especially with large datasets due to the need to test all datasets (Abo-Tabik et al., 2021). Early-stopping can prevent the issue of over-training by imposing constraints on the model parameters.

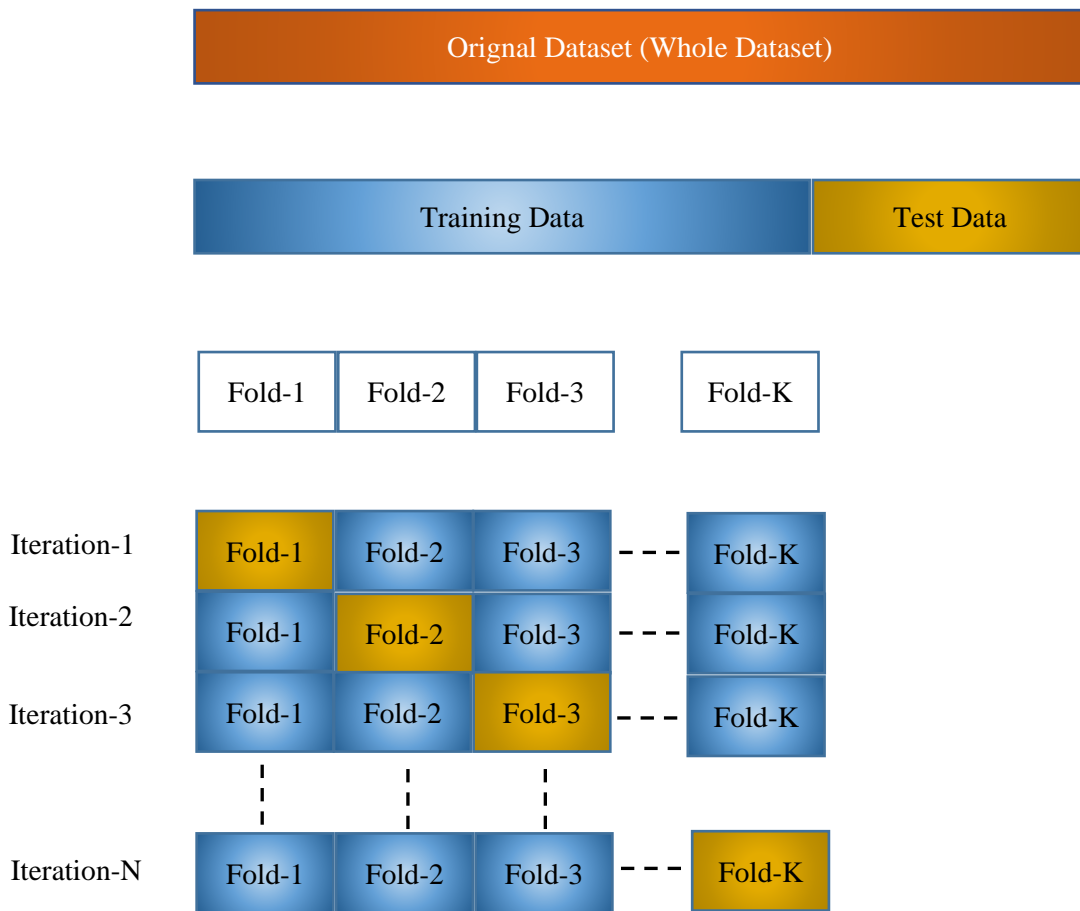


Figure 3.3 K-Fold Cross Validation

Regular evaluation of a model’s performance is crucial to explore its effectiveness at mastering specific tasks and to assess its learning ability in generalizing to unseen data (Liu et al., 2024). The confusion matrix is an evaluation method for multiclass classification tasks that provides better insight into prediction results than performance accuracy metrics. This matrix provides detailed information on true class labels for evaluating the performance of classification models (Helmud et al., 2024).

3.2 Machine Learning Models

The following sections will highlight and briefly describe the most common ML models known for their efficiency in solving classification tasks that have been used in this project.

3.2.1 Support Vector Machine (SVM)

Support vector machine (SVM) is a well-known supervised ML algorithm employed in optimisation methods and statistical learning for solving regression, outlier detection, and classification problems such as FER (Rani et al., 2022). The theory of this model was developed since in the 1990s and proposed in 1996 by Vapnik and other contributors (Vapnik et al., 1996).

The work of SVM is based on finding a hyperplane to separate samples classified with predefined labels, or desired targets by performing optimal data transformations to identify boundaries between data points (Meenal & Selvakumar, 2018). This algorithm has been widely adopted in various disciplines such as healthcare, computer vision, natural language processing, FER, and speech and image recognition. The advantage of this method is memory efficiency due to its need to store only a subset of training points called support vectors to make its decision. In addition, it can perform more effectively than other ML algorithms in high-dimensional feature spaces with small datasets. (Tao et al., 2018; Rani et al., 2022). However, despite its advantages, this model struggles with large training datasets requiring significant computational resources and processing time. Furthermore, it cannot provide an estimate of probability for the prediction.

Moreover, SVM models suffer when dealing with imbalanced data in which one class has significantly more samples than the other classes. The main objective of an SVM model is to determine the hyperplane to separate the data points of various classes. This hyperplane is localized in a way that provides a maximized margin of separation data between classes. Therefore, the highest possible functional margin provides the best data point separation.

SVM models are broadly classified as linear and non-linear SVM classifiers. Linear SVM models can classify linearly separable data by a hyperplane. However, a non-linear SVM needs a kernel function to transform a non-linear discrimination dataset from its dimension to a higher dimension by mapping the data points to space with higher dimensions. For instance, for a dataset in 2D coordinates (x, y) , the kernel function will transfer the data into 3D space (x, y, z) to easily separate and classify data into classes.

The SVM decision function is calculated by the following equations:

$$f(x) = \left(\sum_{i=1}^n \alpha_i y_i x_i \right)^T x + b$$

$$= \sum_{i=1}^n \alpha_i y_i \langle x_i, x \rangle + b \quad (3.6)$$

Where α is the Lagrange multiplier, x input feature vector, $\langle \cdot, \cdot \rangle$ is the inner product operation between two vectors x_i, x , and b is the bias.

Figure 3.4 illustrates the SVM support vectors with a maximised separation margin for a linear separation problem.

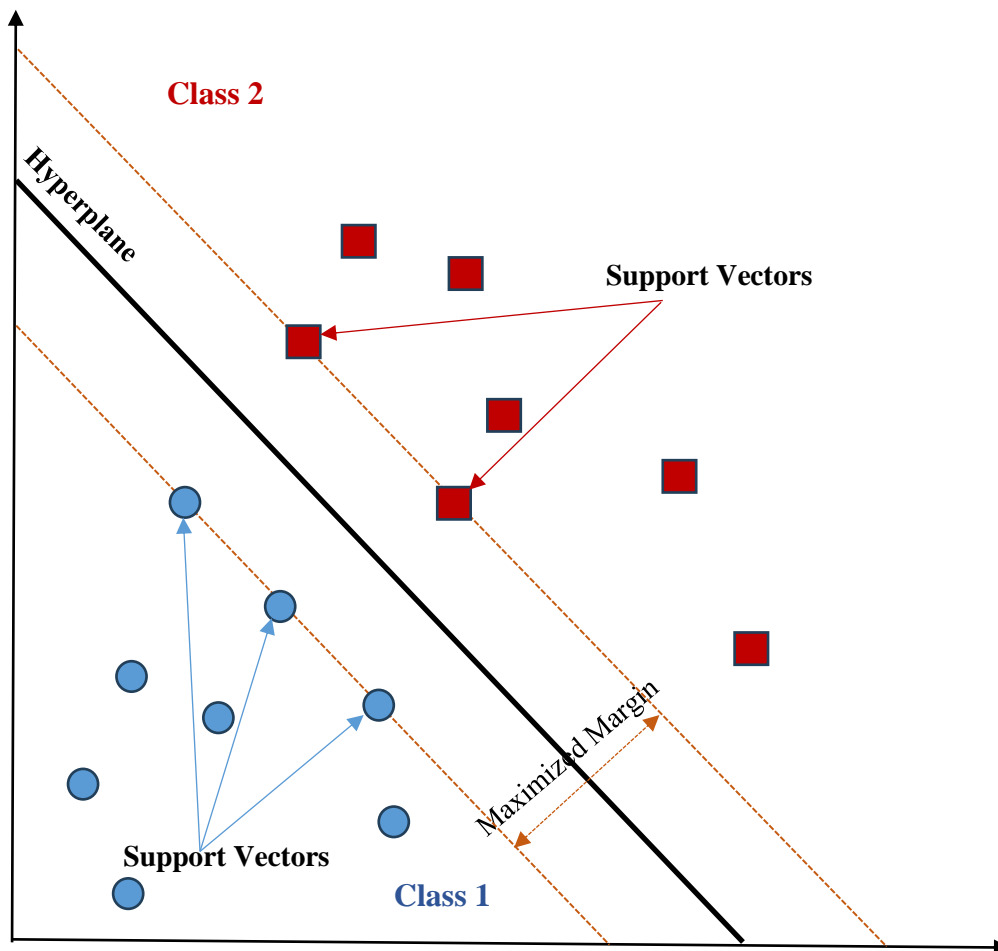


Figure 3.4 An Example showing a Linear SVM Classifier (Wang et al., 2017; Zidi et al., 2018).

However, non-linear discrimination data will need a more complex separation method to map the data points to a higher-dimensional space (see Figure 3.5).

Thus, the detection function is defined as:

$$f(x) = \sum_{i=1}^n \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b \quad (3.7)$$

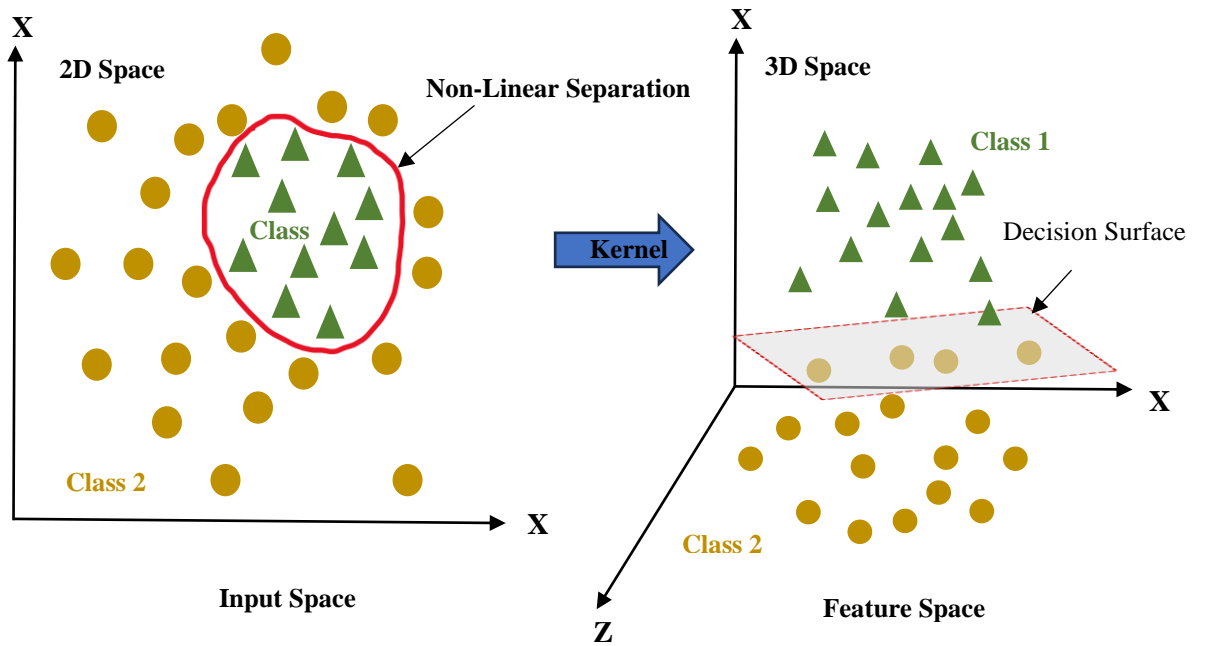


Figure 3.5 Non-Linear SVM Classifier (F. Wang et al., 2017; Zidi et al., 2018).

In higher-dimensional space, a hyperplane is defined as a group of points that have a constant dot product with a given vector. Consequently, the kernel function will take the input vector and produce the dot product of this vector in the feature space.

$$K(x, z) = \langle \phi(x), \phi(z) \rangle \quad (3.8)$$

The utilisation of the low dimensional $K(x_i, x)$ is efficient for computing the inner product in the higher dimensional kernel. The kernel function is calculated as $\langle \phi(x), v(x) \rangle$ (Wang et al., 2017). The accuracy of the classifier is identified by both the functional margin and kernel parameters (Wang et al., 2017; Meng et al., 2019).

3.2.2 Random Forest (RF)

Random Forest (RF) is a powerful well-known ML model for classification and regression problems. The algorithm is based on multiple decision trees and is known for providing higher performance and accuracy than other ML classification algorithms. Each decision tree is grown using a random subset of data. Because of its ability to handle various data, RF is considered suitable for dealing with complicated tasks. In addition, it can operate with imbalanced classes when one class is more infrequent than other classes. Furthermore, this classifier prevents overfitting by employing a randomness operation during the training phase, improving the generalisation ability of the classifier model with unseen data.

In 2018, Jeong and Ko (Jeong & Ko, 2018) reported a study on FER of drivers in real-time. They presented a system based on facial landmarks as a pre-processing method and WFR for classifying driver FE using three databases, an Extended Cohn-Kanade database (CK+), MMI and the Keimyung University Facial Expression of Drivers (KMU-FED) database. In their work, the DLib machine learning library was employed to detect the face and extract spatial features, then the hierarchical WRF classifier was used to classify the FEs. The model showed a good performance similar to that of DL FER approaches, with 92.6% accuracy for the CK+ database and 76.7% for the MMI dataset.

RF works by constructing multiple decision trees during training, providing better voting output for classification or mean average for regression tasks in the individual trees.

Feature Randomness

RF works by constructing multiple decision trees during training and then finding a higher voting output for classification or a mean average for the regression task of the individual trees, as illustrated in Figure 3.6. It is based on the bootstrapping technique, which uses random sampling with replacement. This process involves training each tree on a subset of the original dataset, and selecting a random subset of features (predictors) at each node of the decision tree as candidates for splitting. This introduces randomness into the decision-making process and helps prevent overfitting by reducing the correlation between trees.

Decision Tree Construction

For each bootstrapped sample, a decision tree is constructed using a recursive binary splitting process. At each node, the algorithm selects the best split among the randomly selected features based on criteria such as Gini impurity (for classification) or mean squared error (for regression). The process continues until the tree reaches its maximum height or a minimum number of samples per leaf node.

Voting (Classification) or Averaging (Regression)

Once all trees are constructed, predictions are made by aggregating the individual predictions of each tree. For classification tasks, the mode (most frequent class) of the predictions across all trees is taken as the final prediction. For regression tasks, the mean prediction across all trees is computed.

Ensemble Learning

The final prediction of the RF is an ensemble of predictions from multiple decision trees. Random Forest is considered a versatile and powerful algorithm used in various ML applications due to its advantages which include its ability to handle high-dimensional data, to deal with missing values and outliers, and to provide estimates of feature importance. Furthermore, it achieves better performance than other machine learning classifiers such as SVM and AdaBoost as it is more robust to overfitting and provides better generalisation because of its randomizing approach to feature selection (Ko et al., 2014; Ko et al., 2013). In the training phase, an RF decision tree extracts a subset from the training dataset using the bagging technique, where each tree is randomly grown in a top to down induction, starting with the root node (Pantic et al., 2005b). Figure 3.6 shows RF classifier.

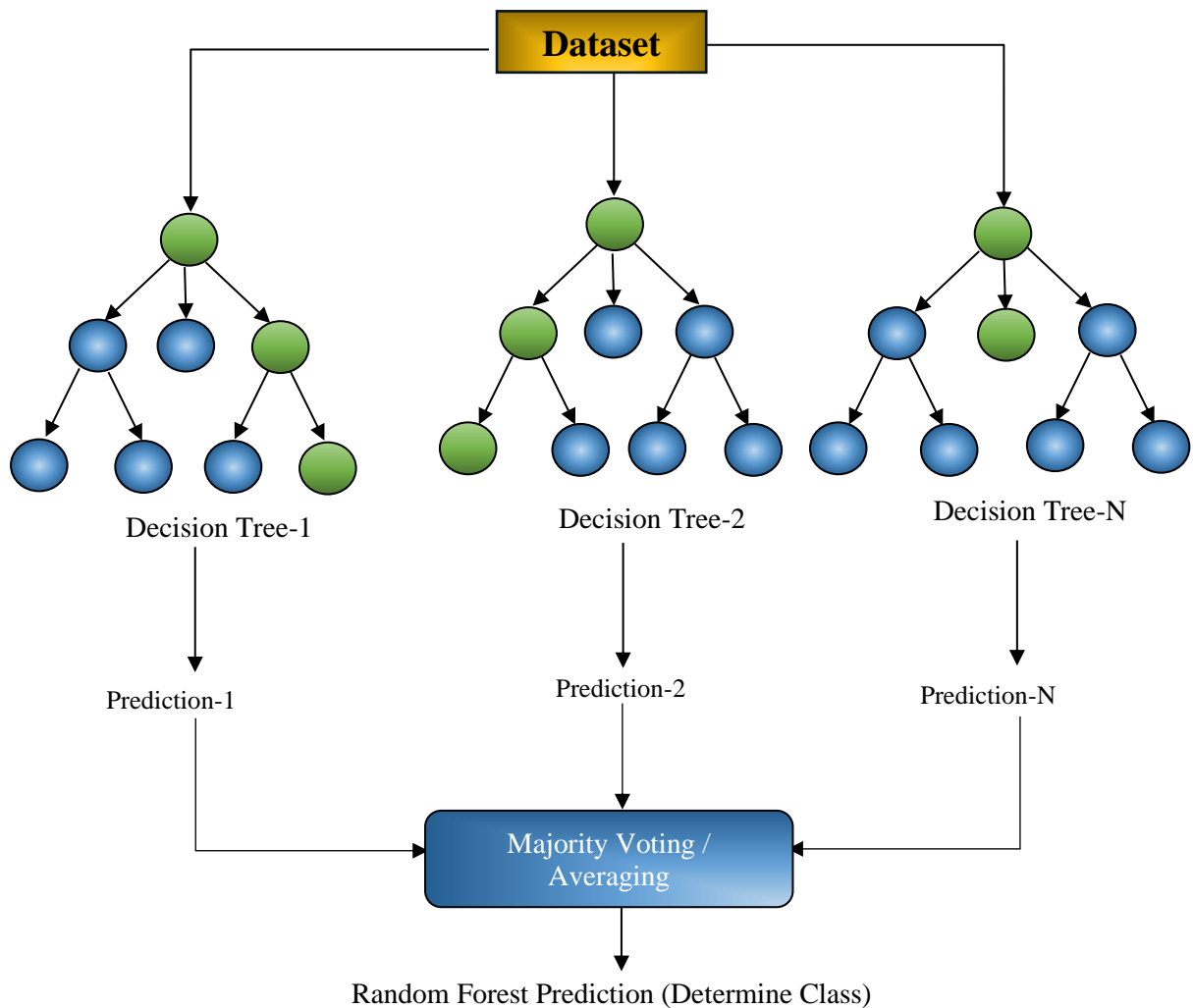


Figure 3.6 Random Forest Model.

3.3 Deep Learning Models

Deep Learning (DL) algorithms are a subfield of ML based on ANN with multiple neurons at multiple layers. They use a form of non-linear mathematical models to process and learn complex representations of data in a method inspired by the human brain. DL algorithms have revolutionised various fields such as computer vision, speech recognition, image recognition, and natural language processing by producing accurate insights and predictions through recognising complex patterns in pictures, text, and sound. In recent years, studies have exploited various versions of DL models to improve the accuracy of prediction (Shi et al., 2015a; Baru et al., 2019).

The following sections highlight three common DL models with good performance in solving various DNN problems, especially those related to images and video data. The third model was designed for natural language processing, but it has also been examined for FER in this thesis. All three DNN model designs along with their evaluation and results will be described in detail in chapter 5.

The first ANN structures were simple fully connected networks known as multi-layer perceptrons (MLP). While MLP can handle many classification and regression tasks, it has poorer performance in tasks with complex data features. The advanced form of ANN is a DL algorithm. The architecture of deep learning models is structured with many layers, each containing multiple neurons. Each layer receives and processes the output data from the preceding layer.

This DL structure and its ability to exploit non-linear features from large datasets increase DL's capability to solve complex tasks. It can capture complex patterns in big data, achieving more accurate results than the older versions of ANN and other ML models.

3.3.1 Improving Facial Expression Recognition via Deep Learning

Recently, along with statistical approaches, many DL methods have been employed in FER applications with remarkable precision. Mehendala (Mehendale, 2020) proposed AFER using convolution neural networks (CNN) in a method based on two main stages. In the first, the background of an image is removed to avoid the negative impacts of external conditions such as distance from the camera. In the second, the facial features are extracted. When this method was applied in databases such Cohn-Kanade expression, Caltech Face Dataset, CMU, and NIST databases, the precision of the proposed AFER model was over 96%.

This thesis aims to investigate the feasibility of developing and constructing a FE's prediction model for patients at risk of deterioration, utilising various deep learning models.

The objective of the model is to alert healthcare professionals to changes in patient health status based on FER and subtle changes in FE.

3.3.2 Techniques to Improve Performance of DL Models

While DL techniques have advanced significantly, recording excellent performance across various tasks, they have limitations due to their complex architectures of large numbers of layers, their complex design, and the migration of processed data across hidden layers. Various techniques have been exploited to enhance performance of DL models and to prevent the loss of essential features during the learning process. The following sections highlight some of these techniques.

The Drop Out Technique

This technique was introduced by Hinton et al. (Hinton et al., 2012) to avoid overfitting and improve the generalization process by randomly discarding neurons based on a specified probability assigned at the design stage and on all the outputs of these nodes. All forward and backward connections with discarded neurons are temporarily removed during the training stage, meaning the drop out is turned off during testing but the dataset will be processed by all nodes including those temporarily discarded. Figure 3.7 illustrates the Drop Out technique.

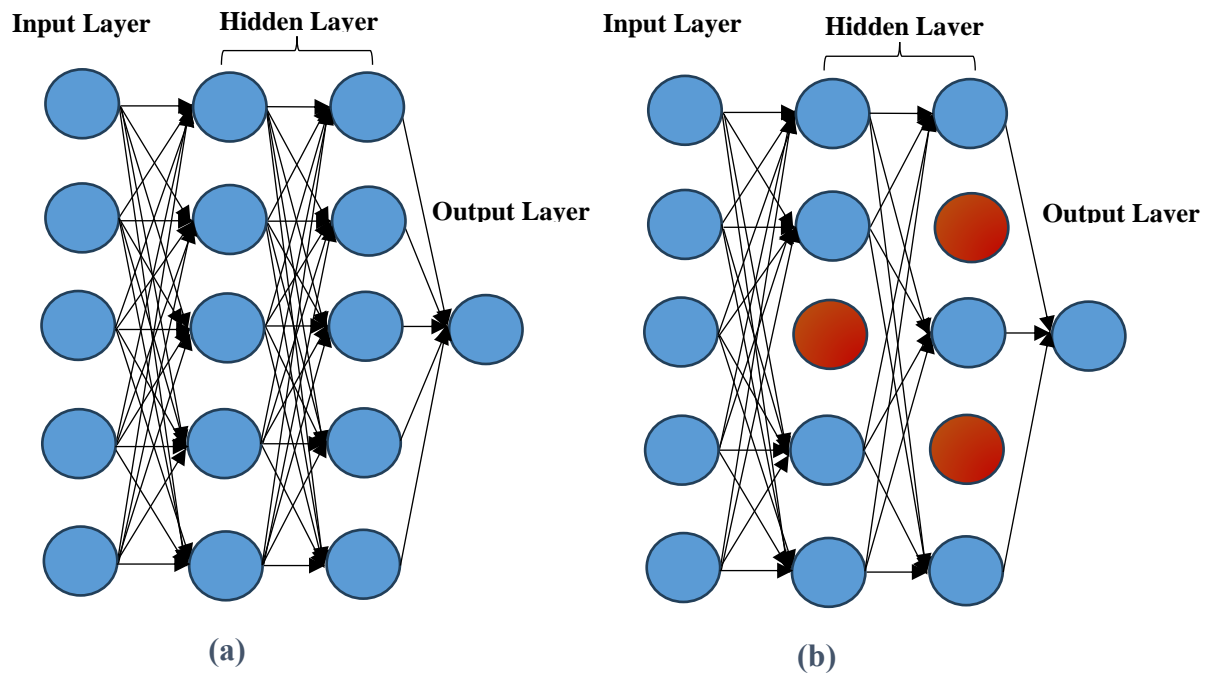


Figure 3.7 Drop out Technique.

(a) Multi-Layer Perceptron Network.

(b) Multi-Layer Perceptron Network after applying dropout.

Batch Normalization Technique

It supports DNNs in accelerating their convergence and improving training stability. However, the different distribution of each input batch, multiple hidden layers, and updated weights during the learning process all contribute to a problem known as internal covariation, which results in a sluggish training process (Ioffe & Szegedy, 2015).

The problem has been solved by batch normalization through minimising internal covariation by fixing the batch inputs for inner layers in the training process. This technique provides maximum stabilisation for deep DNNs through normalising the output of a layer before feeding it to the next layer. It works by subtracting the batch mean, then dividing by the standard deviation according to the following equation for input $x = \{x_1, x_2, x_3, x_4, \dots, x_n\}$:

$$x_i = \frac{x_i - E[x_i]}{\sqrt{Var[x_i]}} \quad (3.9)$$

Where: $E[x_i]$ and $Var[x_i]$ are the mean and variance for each I unit.

3.3.3 The Convolution Neural Network (CNN)

Convolution neural networks (CNN) are a well-known type of DNNs with proven efficiency in pattern recognition tasks with images. They give strong performance and accurate prediction from large non-linear data due to their ability to capture, recognise, and learn complex data patterns through extracting feature vectors (Abo-Tabik et al., 2021). Another advantage is their ability to carry out parallel computations (Sorokin et al., 2018).

CNN was first introduced in 1998 by Yann LeCun et al. (LeCun et al., 1998), and was improved years later by Alex Krizhevsky in collaboration with Llya Sutskever and Geoffrey Hinton (Krizhevsky et al., 2012b; Jogin et al., 2018). LeCun and along with colleagues proposed and developed the basic idea of CNN. The paper of Yann LeCun (LeCun et al., 1998) introduced LeNet-5 architecture, designed and developed at that time for handwritten digit recognition tasks. The architecture of this network consisted of convolution layers and pooling operations, followed by fully connected network layers. The concept work of LeCun inspired the development of modern CNNs which demonstrate high effectiveness in pattern recognition for spatial data.

In 2012, through the invention of AlexNet (Krizhevsky et al., 2012b), CNN was developed and further improved to work on recognizing patterns of images. Despite the need to expand the amount of training dataset, the reason AlexNet has achieved ground-breaking results is the maximizing of learning capacity of this network by adding multiple layers of neurons, and the flexibility of its architecture (Zeiler & Fergus, 2013; Jogin et al., 2018). Another advantage of this model is its use of computational methods on GPUs to process the dataset through the training stage (Jogin et al., 2018).

A convolution neural network (CNN) is an efficient type of DL networks for handling image classification and image analysis tasks due to its specialisation in recognising patterns and complex textures in images. CNN is a sequence of five types of layers including 1) an input layer used for holding the raw data, 2) a convolution layer, which performs convolution, or mathematical operations called dot product between data and filters, 3) pooling layers, which help reduce computational costs, 4) fully connected layer, and 5) an output layer (Jogin et al., 2018).

Convolution layers shift the filter (kernel) across the input data, stopping at each point where a multiplication operation is applied, and the result is summed into the feature map. CNNs use multiple filters to extract different features from the input image, resulting in multiple feature maps, minimizing the need for the complex feature extraction methods usually needed by ML algorithms, and achieving better accuracy by enabling the model to extract its feature map (Fu et al., 2019). The convolution operation can be calculated by the following equation:

$$y = F (X|\Theta) = f_L (... f_2(f_1(X|\Theta_1|\Theta_2)|\Theta_n) \quad (3.10)$$

where n is the number of hidden layers, y is the predicted output, X is the set of inputs, f_1, f_2, \dots, f_L are sequential layers of the CNN (including convolution operations, activation functions, and possibly pooling and fully connected layers), $\Theta_1, \Theta_2, \dots, \Theta_n$ and Θ_i represents parameters (filters/kernels, biases, weights) in each corresponding layer. The convolution operation for layer i is:

$$y_i = f_i (X_i|\Theta_i) = h(W \otimes X_i + b), \quad \Theta_i = [W, b] \quad (3.11)$$

Where the output of the i -th convolutional layer, f_i is the function of the i -th layer applied to the input, the input data to the i -th layer, \otimes represents the convolution operation and W and b are weights and bias respectively (Abdoli et al., 2019).

There are some disadvantages of the feature map produced as an output of convolution operation, particularly its dependence on the location of the important features. Any slight translation in a feature location in the input data between samples results in the production of a new feature map (Kim & Cho, 2018; Kim & Cho, 2019). Therefore, a pooling layer is required after each convolution layer. The pooling layer operation is based on the down-sampling method, where the pooling helps to reduce the spatial dimensions of the feature maps while retaining their important features. Hence, it is used for dimensionality reduction of the convolution layer feature map to enhance the operation of feature extraction and reduce useless computations (Huang et al., 2021; Phan et al., 2016).

Pooling layers are typically inserted between consecutive convolution layers in a CNN architecture, and a final classification layer is usually used. This can be any ML classifier such as SVM, but fully connected network layers are usually used, simplifying the training process (Niu & Suen, 2012; Xue et al., 2016). CNNs with many parameters may suffer from overfitting, especially when the training dataset is small or when the model architecture is too complex. Feature maps with a high capacity to memorise training examples may generalise poorly to unseen data, leading to reduced performance on validation and test datasets. Figure 3.8 shows the general 1D-CNN design.

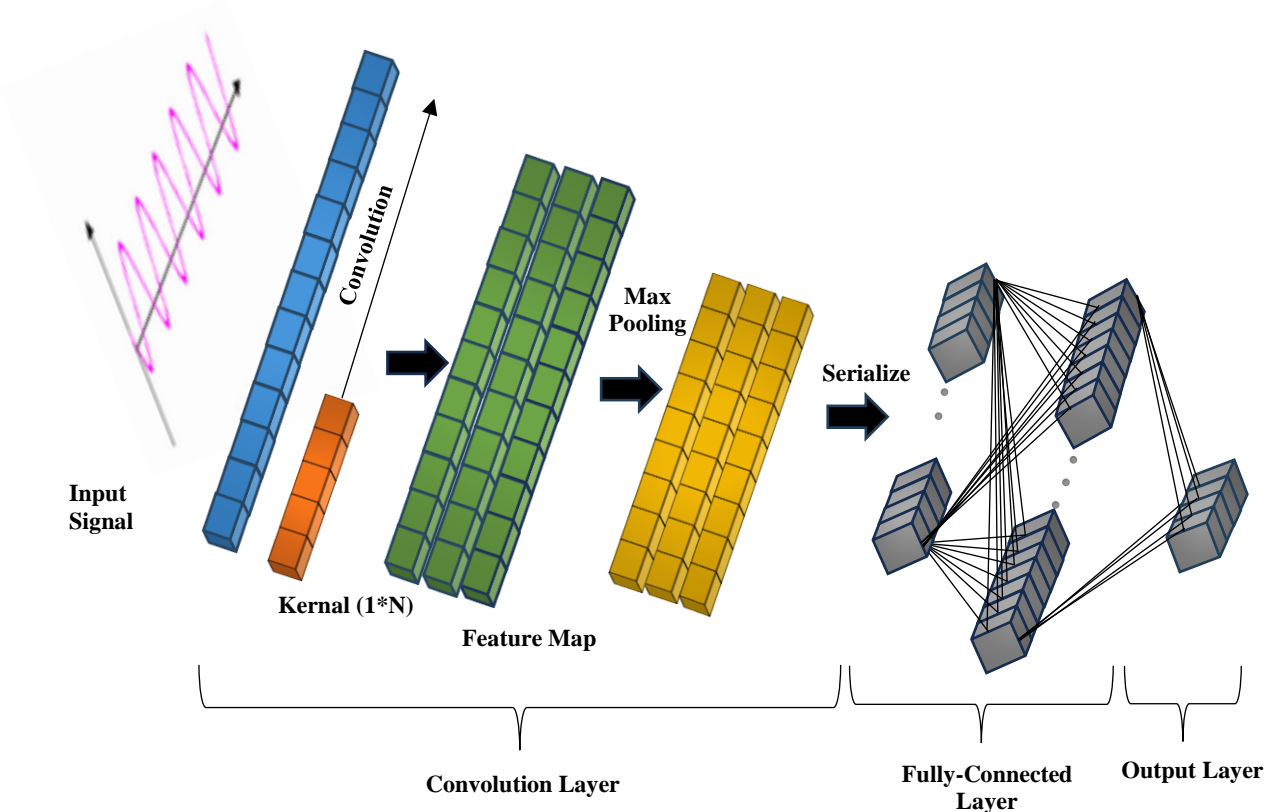


Figure 3.8 General Architecture 1D-CNN based on (Kim et al., 2020).

3.3.4 Recurrent Neural Network and Long Short-Term Memory

A recurrent neural network (RNN) (Rumelhart et al., 1986) is a modified version of a feed-forward ANN, in which the output at each point is linked to all previous inputs. This means that each layer in the current hidden state is a function of the current input and the previous hidden state. For the input sequence $x = [x_1, x_2, x_3, \dots, x_n]$ using RNN, $h_t = f(h_{t-1}, x_t)$ where x_t is the input at time-step t , and h_{t-1} is the previous hidden state (Bouktif et al., 2018). Figure 3.9 shows the general architecture of an RNN.

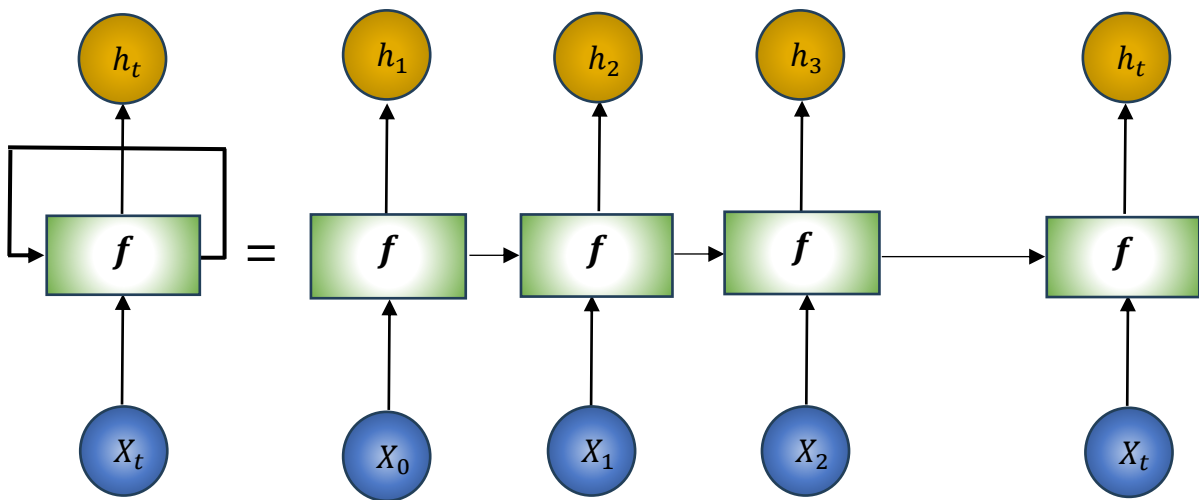


Figure 3.9 Basic architecture of RNN based on (Zhao et al., 2017).

Although this architecture is useful for extracting patterns from time-dependent data samples, it can only look a few steps back because of the vanishing and exploding gradient problem, which can make RNNs challenging to train (Zhao et al., 2017). The long short-term memory (LSTM) is an improved version of RNN designed to overcome long-term dependency problems. This network is trained using a back propagation method (Yan et al., 2018).

Each LSTM layer is a set of blocks which consists of several multiplicative units and memory cells that are recurrently connected. These memory cells are considered the main contribution to the LSTM architecture. The memory cell has three gates and stores the information that is obtained at this step. It then either keeps it, releases it, or resets it according to the state of the controlling gate. The memory unit gates are called input, output, and forget gate, and each is controlled by a sigmoid activation function (Tian et al., 2018).

Figure 3.10 shows the main structure of the LSTM memory unit. Like the RNN, each gate receives an input X for the time t and the previous hidden state h_{t-1} . The forget gate f_t determines how much information will be kept from the previous state C_{t-1} ; f_t is calculated using (Wang, Gan, et al., 2019; Zhou et al., 2016),

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \quad (3.12)$$

where W_f , b_f are the weight and the bias for the forget gate. The input gate i_t , on the other hand, is responsible for controlling the amount of current information to be considered as input for generating the current state C_t ; i_t is calculated using equation (3.13); W_i , b_i are the weight and the bias for the input gate (Wang, et al., 2019; X. Zhou et al., 2016).

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (3.13)$$

Now the current hidden state C_t will be calculated using the long-term information obtained from f_t and the short-term information from i_t , as in the following equations, W_c , b_c , are the weight and the bias for the current state (Wang et al., 2019; Zhou et al., 2015).

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3.14)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (3.15)$$

where $\tanh(\cdot)$ is the activation function, and $*$ is an element-wise product. The last gate o_t is the output gate, which will decide the amount of information to be treated as output, and will be calculated using:

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (3.16)$$

Where W_o , b_o are the weight and the bias for the output gate. Finally, since all gates are controlling the information flow using the element-wise product; the final out h_t will be calculated (Wang et al., 2019; Zhou et al., 2015) as

$$h_t = o_t * \tanh(c_t) \quad (3.17)$$

LSTM algorithms have achieved acceptable performance in FER by measuring changes in movement between the previous, current, and next frames of a video stream.

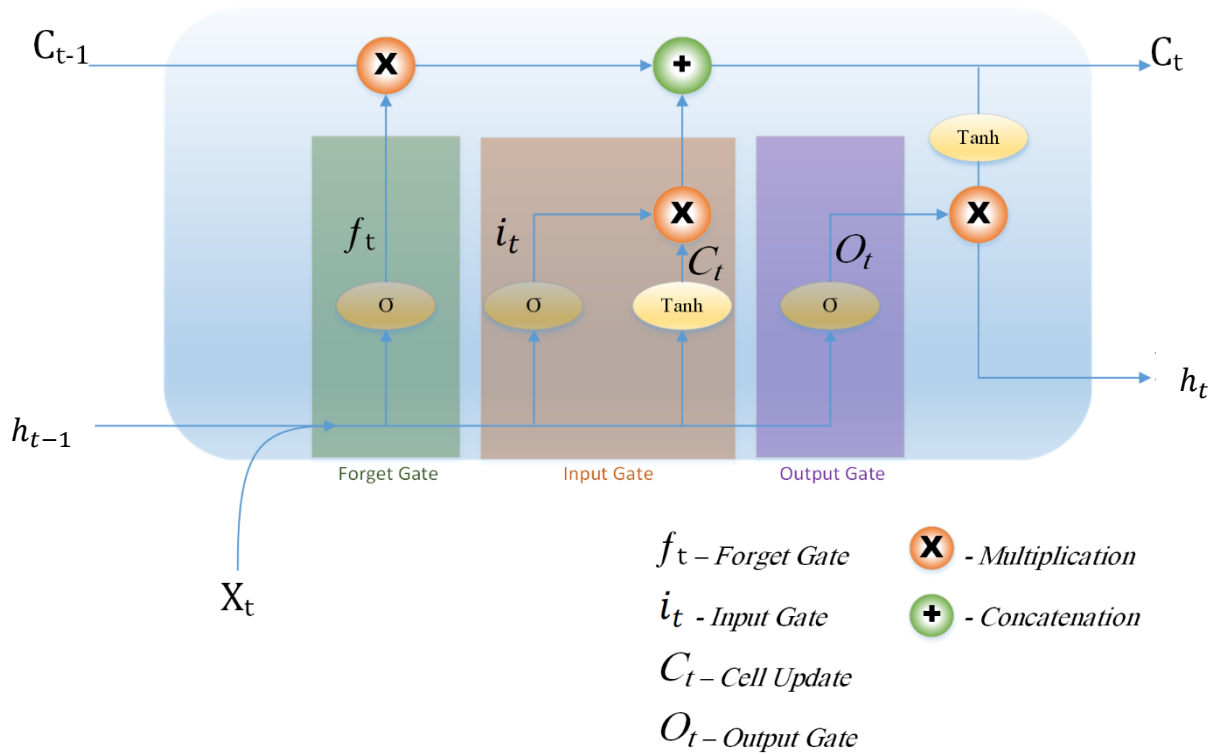


Figure 3.10 Inner structure of LSTM cell (Shi et al., 2022).

3.3.5 Convolution LSTM

Data collected over successive periods of time are characterised as a time series. In such cases, an interesting approach is to use a model based on LSTM, a RNN architecture. In this kind of architecture, the model passes the previous hidden state to the next step of the sequence. Therefore, the network holds information on previous data it has seen before and uses it to make decisions. In other words, the data order is extremely important over time and the features of this type of data are known as temporal features. For image analysis and classification tasks, CNNs are the adopted approach. The image passes through convolutional layers in which filters extract features of interest known as spatial features. After passing various convolution layers in sequence, the output is connected to a fully connected dense network.

Spatio-temporal prediction is challenging due to complex dynamics and appearance changes which impose dependencies on both temporal and spatial domains. In AI-based methods, integrating multiple DL models can produce highly accurate predictions and improve model performance (Moishin et al., 2021). A common effective combination model for detecting and recognising complex features within sequences of images is CNN with LSTM, known as ConvLSTM. This model offers high predicting accuracy and performance in capturing spatial and temporal features through video stream (Miyoshi et al., 2021). ConvLSTM model has

been proposed by group of researchers (Shi et al., 2015b), where it proposed as a variant of the traditional LSTM involving convolution operations into the LSTM cells. The ConvLSTM model was specifically designed for spatio-temporal sequence prediction tasks, such as precipitation nowcasting in which both spatial and temporal dependencies are important.

The convolution operations in ConvLSTM allow the model to capture spatial patterns in the input data, while the LSTM cells enable it to capture temporal dependencies over time. ConvLSTM replaces all its linear operations with convolution layers to capture spatial dependencies besides the LSTM, and many of its variants (Wang et al., 2017; Wang et al., 2018; Zhang et al., 2019) have achieved impressive results in spatio-temporal prediction. Therefore, ConvLSTM has been widely adopted in various applications, including weather forecasting, video processing, and medical imaging, where spatio-temporal data analysis is crucial. Figure 3.11 illustrates the ConvLSTM structure (Zhang et al., 2018) where the new memory C_t and output H_t will be generated by updating the internal memory C_{t-1} to the current input X_t and the previous output H_{t-1} .

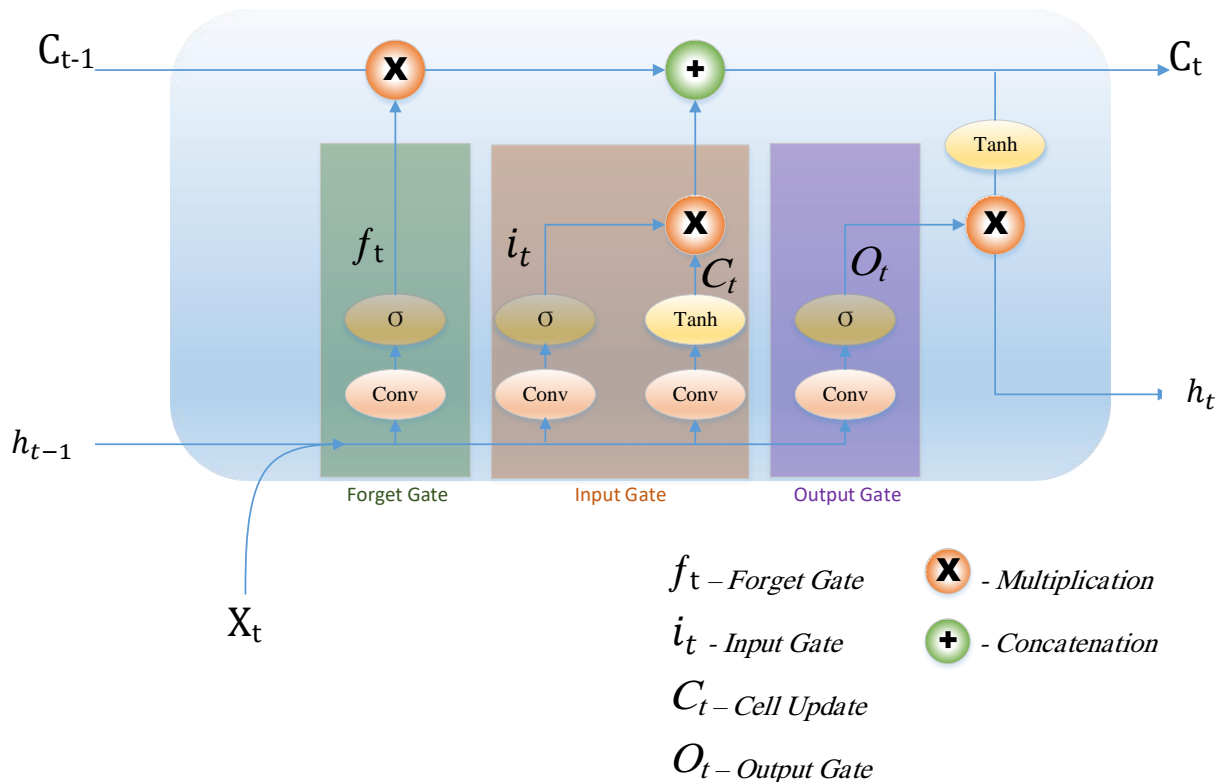


Figure 3.11 The structure of ConvLSTM. The new memory C_t and output h_t will be generated by updating the internal memory C_{t-1} according to the current input X_t and the previous output h_{t-1} (Zhang et al., 2019).

The mathematical expression of the ConvLSTM in the updated gates is given as follows:

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * h_{t-1} + W_{cf} * C_{t-1} + b_f) \quad (3.18)$$

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * h_{t-1} + W_{ci} * C_{t-1} + b_i) \quad (3.19)$$

$$O_t = \sigma(W_{x0} * X_t + W_{h0} * h_{t-1} + W_{c0} * C_t + b_0) \quad (3.20)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tanh(W_{xc} * X_t + W_{hc} * h_{t-1} + b_c) \quad (3.21)$$

$$h_t = O_t \times \tanh(C_t) \quad (3.22)$$

Where $*$ refers to convolution operation, \times refers to the Hadamard product, and W_{cf} , W_{ci} , W_{c0} refers to the weight matrices. All the weight matrices and bias vectors will be updated in each update process. If we view the states as the hidden representations of moving objects, a ConvLSTM with a larger transitional kernel should be able to capture faster motions while one with a smaller kernel can capture slower motions (Shi et al., 2015a).

3.3.6 Transformers

The natural language processing (NLP) field witnessed a revolution on the introduction of a new type of DL algorithm known as Transformer, which was first proposed in 2017 by Vaswani et al. (Vaswani et al., 2017). They presented a simple network architecture, the original transformer, that uses a self-attention mechanism to capture long-range dependencies within sequences, eliminating the need for RNNs or CNNs. Transformers have become a cornerstone architecture in natural language processing (NLP) and have been widely adopted in various other domains, including computer vision. Therefore, some researchers recently started to be interested in involving Transformers into FER systems. Several recent studies have been reported, including one (Yuan et al., 2022) addressing the problem of a face mask in pain assessment, recognising pain intensity from just the upper part of a face. They proposed a Swin-Transformer model conducted on the UNBC-McMaster database which includes data related to facial AUs that are typically associated with pain expression. The model achieved accuracies higher than 90% with and without face mask.

In the same year, another group of researchers applied the Vision Transformer model to FER using the Indonesian Mixed Emotion Dataset (IMED) and the of the model achieved outstanding performance (Akeh et al., 2022). One year later, a study by (Vats & Chadha, 2023) presented an AFER framework capable of recognising seven FEs with minimal dataset using Swin Vision Transformers (SwinT), squeeze and excitation (SE), and spatial attention

module (SAM) to optimise the performance of the model. The accuracy of the model was around 53%. In the same year, a FER framework introduced by Safavi, Patel, and Vinjamuri (Safavi et al., 2023) involved a hybrid system of Mix Transformer and an additional fusion block. The model employed a self-attention technique to effectively concentrate on pixel-level landmark segments. The model showed a good accuracy around 73% when conducted on the FER2013 database.

Different types of transformer models and architectures exist.

1. Original transformer

Presented in the paper "Attention is All You Need" by Vaswani and other contributors (Vaswani et al., 2017), the original transformer model features an encoder-decoder architecture. It uses self-attention mechanisms to process input data in parallel and transformer blocks that consist of multi-head attention and position-wise fully connected layers.

2. Bidirectional encoder representations from transformers (BERT)

Developed by (Devlin et al., 2018), BERT revolutionised the understanding of context in language by reading input data bidirectionally. This model is primarily used as a foundation for a range of NLP tasks, demonstrating remarkable performance across various benchmarks.

3. Generative pre-trained transformer (GPT)

GPT models are a series of language models developed by OpenAI (Radford et al., 2018) based on the transformer architecture and optimised for natural language understanding and generation. Each version of GPT (from GPT-1 to GPT-5 and beyond) has increased in complexity and capability, featuring more layers and parameters for deeper contextual understanding, advancing the field of natural language processing (NLP).

3. Transformer-XL

In 2018, (Dai et al., 2018) introduced a variant Transformer architecture designed for natural language processing (NLP) tasks. It reuses the hidden states from previous segments, allowing them to learn dependencies beyond a fixed length without disrupting temporal coherence. It is particularly effective in tasks requiring long-range memory, such as document classification and question answering.

5. XLNet

Developed by Google and Carnegie Mellon University (Yang et al., 2019), XLNet integrates aspects of BERT, Transformer-XL, and state-of-the-art autoregressive models. It represents a significant evolution by improving the limitations of the

BERT architecture, especially in handling sequential data for tasks such as question answering and document classification, as it uses a permutation-based training method rather than the traditional masked language model, allowing it to learn bidirectional contexts and outperform BERT in many benchmarks.

6. T5 (Text-To-Text Transfer Transformer)

The T5 model developed by Google (Raffel et al., 2020) frames all text-based language tasks into a unified text-to-text format, where every task is cast providing some input text with an expected output text. This simplifies the use of a single model for diverse tasks like translation, summarisation, and question answering, among others.

7. Vision Transformers (ViT)

ViT is an adaptation of the transformer architecture for image classification tasks and introduced by a team of researchers at Google Research (Dosovitskiy et al., 2020), where the image is split into patches and the sequence of linear embeddings of these patches is processed by a transformer. ViT demonstrates that transformers can be directly applied to images without convolution layers, achieving state-of-the-art results in many vision benchmarks.

8. DETR

DETR introduced by a team of researchers from Facebook AI Research (Carion et al., 2020) uses the transformer's self-attention mechanism to eliminate the need for many manual components in object detection systems, providing a streamlined architecture for both object detection and instance segmentation tasks.

9. RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa is a variant of BERT introduced by researchers from Facebook AI and the University of Washington (Liu et al., 2019) that modifies key hyperparameters and removes the next sentence pretraining objective. RoBERTa trains longer with more data and larger batches, achieving improved results over BERT.

10. ALBERT (A Lite BERT)

ALBERT is a simplified version of BERT introduced by Google Research and Toyota Technological Institute at Chicago (Lan et al., 2019) that reduces model size (but not necessarily complexity or training time) by factorising the embedding layer and sharing parameters across hidden layers.

Each of these transformers has its specialties and optimisations designed to tackle specific problems or improve efficiency and effectiveness in learning representations from data. The choice of a transformer model often depends on the specific task, the size and nature of the

dataset, and the computational resources available. As the nature of the generated data are images and sequences of frames, the model proposed in this thesis was Vision Transformer due to its adaptivity to this type of data and the task target (FER). This model was presented by (Dosovitskiy et al., 2020) based on a pure transformer that implemented directly sequences of image patches without the reliance on CNN, attaining highly accurate results.

ViT treats an image much like a sequence of words (tokens) in a sentence, performing the Transformer architecture directly on patches of the image. In addition, ViT cannot recognise the order of the input, therefore positional embeddings are added to the patch embeddings to retain positional data. The overall structure of ViT architecture consists of several stages. Firstly, the input image splits into fixed-sized patches (e.g., 18*18) using the following formula:

$$N = \frac{HW}{P^2} \quad (3.23)$$

Where:

N is the number of patches.

H is the height of the image.

W is the width of the image.

P^2 each patch of size ($P * P$) pixels.

These patches are flattened and linearly transformed to a lower dimension (D), including positional embeddings to create a sequence of vectors (one vector for each patch). Then, the sequences are fed to the transformer encoder with image labels. Finally, the output of the last transformer block is passed through a classification head which typically consists of a single fully connected layer. Figure 3.12 illustrates the Visual Transformer.

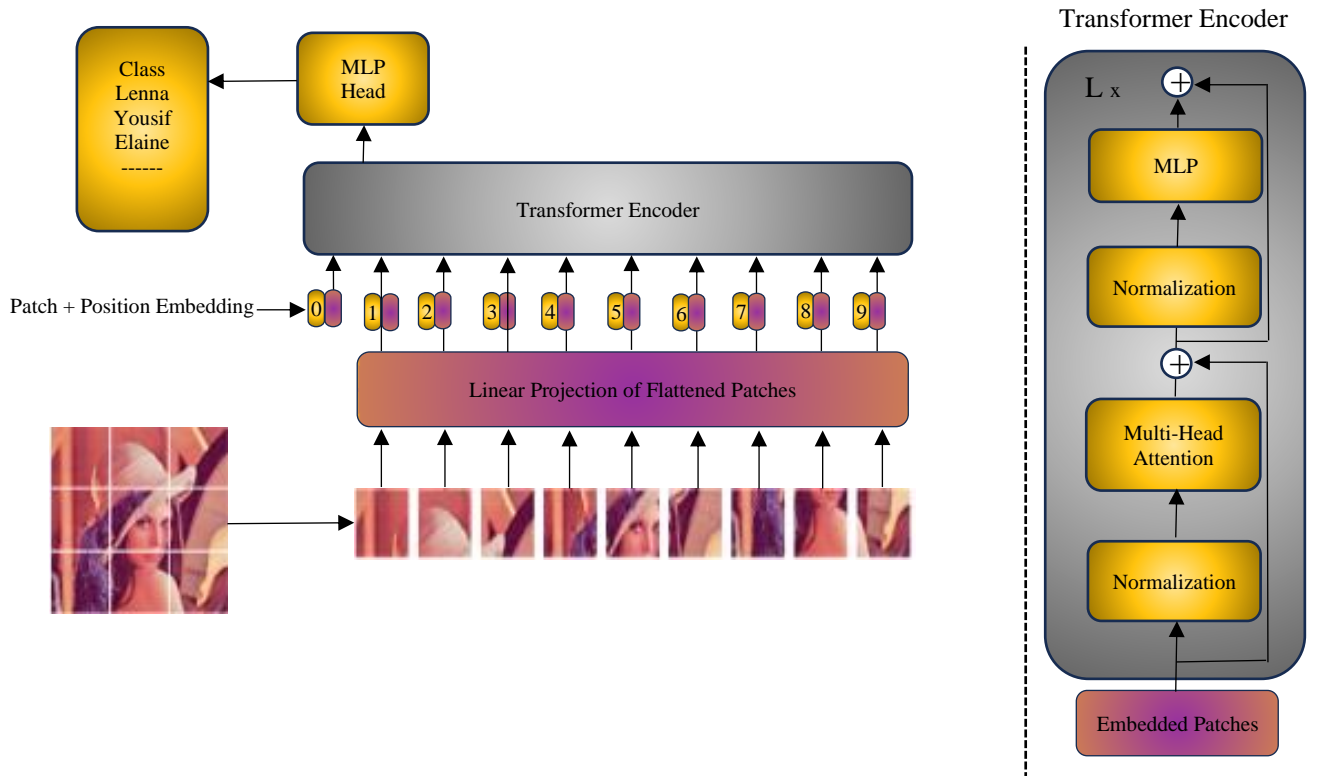


Figure 3.12 Vision Transformer Architecture (Dosovitskiy et al., 2020).

3.4 Some Important DL Model Activation Functions

An activation function is a mathematical function considered an essential part of the DNN architecture. It is a kind of a filter that determines the output of a computation node, and therefore has a significant impact on performance and accuracy of the model (Abo-Tabik et al., 2021). The Softmax activation function (Huang et al., 2018; Tang, 2013) is often used in the output layer of DNN, especially for classification tasks. The Softmax function assigns a discrete probability distribution for K classes which is defined mathematically for input data $z = [z_1, z_2, z_3, \dots, z_k]$ by the following formula:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}} \quad (3.23)$$

Where:

σ is the Sofmax function,

z_i input data,

k number of classes,

e^{z_i} exponential function for input data,

e^{z_j} exponential function for output data.

The rectified linear unit (ReLU) is another activation function which will be widely used in this thesis, and it is one of the most widely used in NN (Agarap, 2018; Zhang et al., 2019; Zou et al., 2020). The ReLU function $f(x)$ is a linear activation function which will return 0 if it receives any negative input and leaves positive values unchanged. Hence, the outcomes of this function fall in the range from 0 to infinity.

$$f(x) = \max(0, x) \quad (3.24)$$

Where:

$f(x)$ the ReLU function, x is the input data.

The selection of activation functions has a significant impact on the performance of NN, so these are defined under experimentation and fine-tuned throughout model development.

3.5 Genetic Algorithm (GA) for Hyperparameters Selection

DL models have a set of hyperparameters which vary for different models and tasks (Abo-Tabik et al., 2021). For instance, each architecture of an LSTM model has a different number of layers, different fully connected network and filter size, different activation functions, etc. Finding the optimum values for each one of these parameters is a time-consuming task for designers and developers who use trial and error methods to search an unlimited number of combinations (Hutter et al., 2019). The Genetic Algorithm (GA) was first introduced by (Fraser, 1957), and later developed by Holland (Holland, 1975). It is an optimization method based on the concept of natural selection and genetics. It uses the concept of “Survival of the fittest” to obtain the optimal values (Meng et al., 2019).

During the initialization phase of GA, a set known as a population refers to a set of randomly generated candidate solutions. Each member in the population is known as a chromosome which represents a potential solution to the optimisation task. The fitness of each member in a population is evaluated with a fitness function to examine how well the member performs with respect to the optimisation target. The fitness function assigns a numerical value to each member, indicating its quality or suitability as a solution. Individuals are selected from the population to be parents for the next generation. Selection is typically based on a selected

fitness function, with higher-fitness individuals having a greater chance of selection (P. Tao et al., 2018). Mutation and crossover operators are applied to the population set to maintain population diversity and prevent local optima. Selected parents undergoing crossover are replaced and combined to produce offspring for the next generation. Crossover promotes exploration of the solution space by combining promising features from different parents.

In contrast, mutation is the process of introducing random changes to individual chromosomal genes in a parent chromosome to generate a new individual offspring. The offspring generated through crossover and mutation replace individuals in the current population to form the next generation. The replacement process ensures that the population size remains constant over generations and that only the fittest individuals survive to the next generation. The evolutionary process continues iteratively through multiple generations until a termination condition is met. Termination conditions may include reaching a maximum number of generations, achieving a satisfactory level of solution quality, or reaching a predefined computational budget.

Genetic algorithms aim to converge towards optimal or near-optimal solutions over successive generations. Convergence is achieved when the population converges to a set of high-quality solutions that satisfy the optimization objective. Overall, genetic algorithms are versatile and can be applied to a wide range of optimisation problems, including function optimisation, parameter tuning, scheduling, and machine learning. They are considered a good solution for introducing better models due to their robustness, flexibility, and their ability to handle complex and non-linear search spaces. However, they may require careful parameter tuning and can be computationally expensive for large-scale problems. GA can improve the search process for better performance and generalization of DL models, and it has been employed for hyperparameter optimisation in ML models (Meng et al., 2019; Sukawattanavijit et al., 2017; Tao et al., 2018; Tao et al., 2019). It has also been employed to select the best values for model architecture in DL tasks. (Bouktif et al., 2018; Bouktif et al., 2020; Lu et al., 2020; Sun et al., 2020).

3.6 Conclusions

In this chapter, various ML and DL models and techniques for improving their performance have been reviewed in detail and will be employed later in the solution proposed in this thesis. The chapter began with a general explanation of ML and what validation and evaluation approaches are commonly used.

Support Vector Machine and Random Forest were then introduced, two ML models which are well-known for classification tasks. Additionally, relevant DL models were introduced, including 1D CNN, ConvLSTM, transformers and techniques that have been introduced to improve the models' performance. Both 1D-CNN and LSTM methods were also reviewed and will be employed to predict risk of deterioration in this thesis. The chapter ends with a brief description of genetic algorithms, which will be exploited for hyperparameter selection and fine-tuning.

Chapter 4

Datasets Description and Analysis

4.1 Introduction

In the healthcare domain, data privacy, availability, and diversity can be significant concerns, especially for training DNN models which require a large amount of diverse data (Rieke et al., 2020). In addition, collecting real samples from patients is a costly and time-consuming process due to procedural and ethical considerations and restrictions to protect patient privacy (Price & Cohen, 2019; Antunes et al., 2022).

The creation of a synthetic database with the latest advanced techniques in computer-generated imagery (CGI) to generate a realistic avatar capable of expressing the full spectrum of FEs is now possible (Safavi et al., 2023). 3D modelling software offers scripting tools for modelling, rigging, animation, simulation, rendering, compositing, and motion tracking. The significance of avatars in expressing FEs has drawn attention for the study of social cognition through HCI (Bombari et al., 2015; Wykowska et al., 2016). However, synthetic data may suffer from uncanny valley, a phenomenon proposed by Mori (Mori, 1970), which describes a lack of realism particularly when the avatars perform the subtle natural movements in FEs (Tinwell et al., 2010). A reason for this failure to reproduce realistic facial expressions may be lack of knowledge about the timeline of synchronised FEs and/or the crucial information perceived from the movement of facial muscles. Therefore, it is important to address and mitigate this issue to generate realistic interactive FEs in the context of health deterioration by investigating the empathy elicited by facial expressions, and their duration (Safavi et al., 2023).

Creating a dataset of animated facial expressions through avatars involves a unique set of steps, focusing on digital creation and animation rather than capturing real human FEs. The process is outlined below and systematically represented in the flowchart (refer to Figure 4.1) that outlines the process for generating such a dataset.

1. Define objectives
 - a. Define the purpose and the specific goals of the dataset, which for our project involves FER to determine five types of FE that indicate a risk of deterioration.

- b. Define data requirements to determine types of FE (and the AUs that comprise them), the diversity of participants (age, gender, skin tone, ethnicity), and environmental conditions (lighting, background).
2. Design avatars
 - a. Create 3D avatar base models by design or selection using 3D modeling software (e.g., Blender, Maya, FacsHuman).
 - b. Customise avatar features to ensure diversity in facial features and FEs (age, gender, skin tone, ethnicity).
3. Animate expressions
 - a. Animate the predetermined FEs using keyframe animation or motion capture technology to create smooth transitions and realistic actions.
 - b. Review animations to ensure that each animation smoothly and accurately conveys the intended FE, making necessary adjustments to animations based on feedback.
4. Data annotation
 - a. Label each animation frame or video with the corresponding FE.
 - b. Add metadata to each frame or sequence, detailing the FE type and avatar characteristics.
5. Rendering

Render the animations to ensure consistent combination of AU appearances, introducing high-quality images or video frames at a defined frame rate (e.g., 20, 25, 30, or 60 fps).
6. Preprocessing
 - a. Apply face detection algorithms to crop images, removing unnecessary background.
 - b. Normalisation to standardise image sizes and colour schemes for uniformity, adjusting the range of pixel intensity values, typically [0, 255]. This is particularly important when images are obtained under different lighting conditions.
 - c. Augmentation of data, applying techniques such as rotation, flipping, and scaling to expand dataset size and improve model robustness. For the proposed model, data augmentation techniques include changes in lighting, and slight modifications in expressions.

7. Review

- a. Perform spot checks on the dataset to ensure all images are correctly processed and labelled.
- b. Correct errors in labelling or poorly processed images.

8. Data splitting

- a. Partition the dataset into training, validation, and testing sets to train, validate, and test ML and DL algorithms.

9. Data storage and management

- a. Save the data in an organised format (pickle file), within a database or file system for easy retrieval.
- b. Backup data and implement recovery protocols to prevent loss.

10. Documentation

- a. Document the dataset creation process, data structure, and access instructions.
- b. Document ethical considerations, especially if the avatars are based on real individuals.

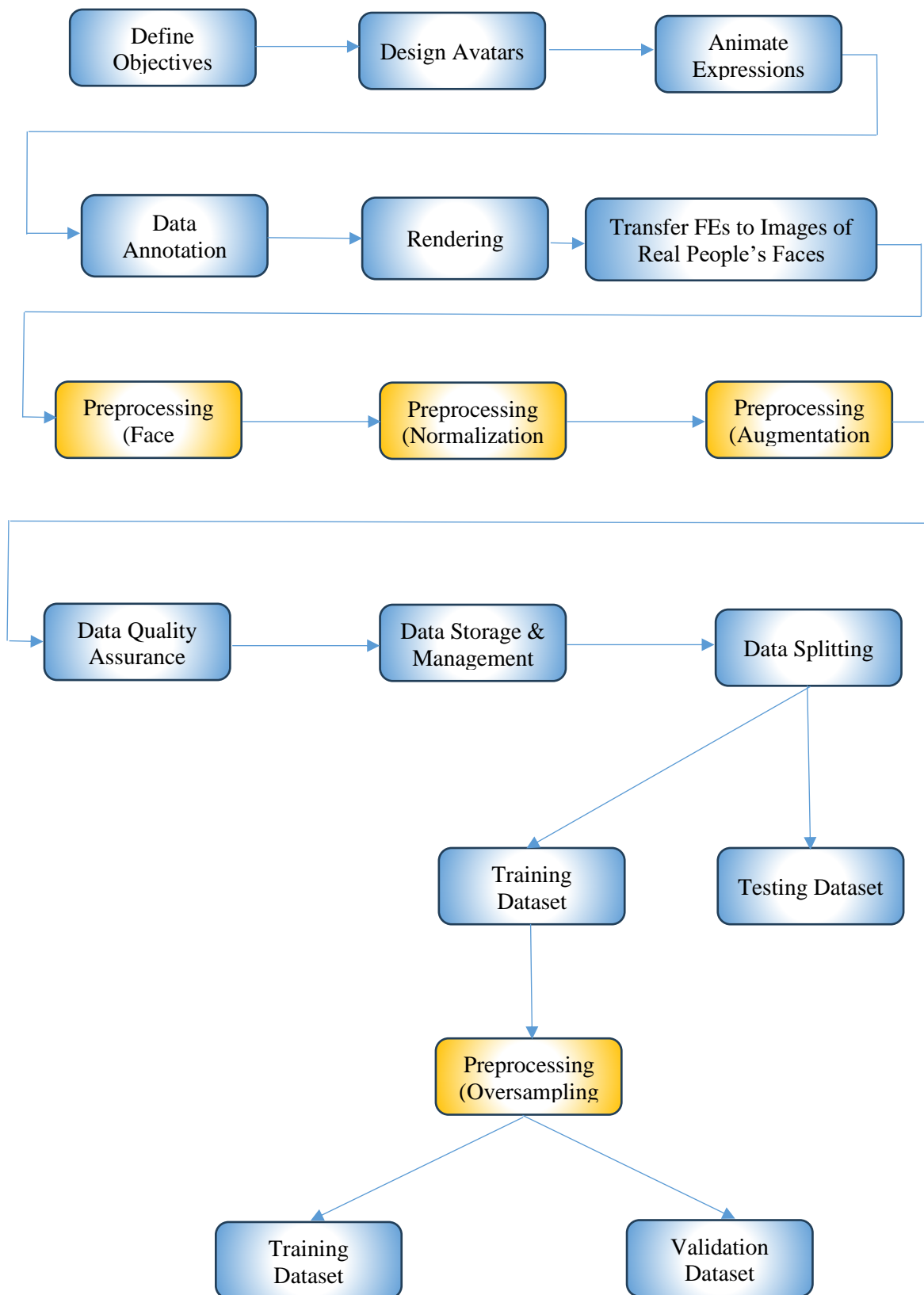


Figure 4.1 Methodology of generating dataset (PRD-FE).

4.2 AUs Analysis

The Facial Action Coding System (FACS) presents sets of action units (AUs) that comprise FEs through changes in facial muscle movement. AUs are a cornerstone for addressing ambiguity in FE analysis, and their identification is crucial for enhancing FER performance. Consequently, this thesis identifies specific AUs that are considered the key discriminatory features for classifying the five classes of FEs that represent deterioration.

Facial AUs analysis involves the detection and interpretation of facial muscle movements, or AUs, to infer underlying emotions or expressions. Each AU corresponds to a distinct facial muscle or group of muscles and is associated with a particular movement.

Analysing AUs can effectively address ambiguity and complexity, achieving an accurate representation of individual expressions and enhancing FER performance. The AU analysis process typically consists of several steps (X. Niu et al., 2019):

The process of AU analysis typically involves the following steps.

1. Facial landmark detection

Detecting key facial landmarks, such as the corners of the eyes, nose, and mouth, which serve as reference points.

2. AU detection

Identifying the occurrence and intensity of specific AUs by analysing the spatial relationships and movements of facial landmarks. This can involve techniques such as template matching, ML classifiers, or DL models trained specifically for AU detection.

3. AU interpretation

Interpreting the detected AUs in the context of known FE patterns or emotion models. This step involves mapping combinations of AUs to FEs.

4. Emotion inference

Inferring the emotions which underlie the detected AUs and FEs. This may involve referencing established emotion models like FACS or using ML algorithms to classify emotions based on AU configurations.

Facial AU analysis is widely used in fields such as psychology, human-computer interaction, and affective computing to study emotions, assess psychological states, and to develop applications for emotion recognition and expression synthesis. However, FE interpretation is ambiguous (Davison et al., 2018), such as the inner brow raiser may refer to surprise or sad. The FACS (Ekman & Friesen, 1978) has been verified to be effective for resolving this ambiguity. In FACS, AUs are defined as the basic facial movements which combine to form multiple FEs (Ekman & Friesen, 1978). The criteria for AU and FE correspondence are defined in the FACS manual, which provides a framework for encoding FER (Xie et al., 2020; Sun et al., 2022; Chen & Joo, 2021) through embedding AU features. Video-based FER methods are important in disciplines (Liang et al., 2023) such as clinical diagnosis and interrogation, allowing identification of micro FEs from video streams. However, this task is a challenge due to the scarcity of datasets (Zhang et al., 2021), and the involuntary, fleet, and low-intensity expressions cannot be mimicked by people. This thesis presents an AFER for patients in deterioration, so the generated data should represent various medical conditions. The data could not be collected from the clinical setting, so the dataset has been generated using avatars or 3D face models of patients in stable and deteriorating condition.

Chapter 2 explored macro and micro expressions and the related AUs in terms of patient health status. If data were to be collected from human participants, facial expressions can be gathered through two approaches, namely posed and spontaneous. A posed database is generated using a supervised method with participants who intentionally present a MaFE. Spontaneous facial expressions are collected from participants who, without prior knowledge of the expected expressions, are presented with a set of stimuli.

The type of database for the proposed project is a spontaneous database that cannot be mimicked by other people due to some facial expressions being involuntary and therefore, cannot be controlled by humans. It becomes important thus, to find an automated supervised method to produce these types of expressions.

This thesis proposes a spontaneous database based on involuntary FEs that cannot be mimicked. Therefore, it is important to identify an automated supervised method to produce these types of expressions. Avatars can produce voluntary and involuntary FEs through controlling AUs and their intensity. These FEs can then be transferred to face images of real people. This solution allows us to test the designed methods on realistic datasets which mimic real samples of patient data, simulating FEs for the medical conditions of interest. Subsequently, a trend analysis will indicate whether the patient is in a stable or deteriorating condition.

The recent published papers reviewed in Chapter 2 have recorded high accuracy in recognizing the seven universal facial macro expressions of emotion categorised as neutral, happiness, sadness, fear, anger, disgust, and contempt. This thesis proposes to use AI tools to detect micro facial muscle movements to evaluate patient health trends. Deep learning algorithms have proved ineffective for this task, having difficulty recognising FEs in the offset stage of change from stability to deterioration.



Figure 4.2 Facial expression at different stages

Figure 4.2 illustrates a sequence of patient FEs from stable through to critical deterioration stage. Note that critical deterioration is characterised mostly by a combination of the eyes, lips and head pose.

Based on the study of (Madrigal-Garcia et al., 2018) whose stable and deteriorated cases are illustrated in Figure 4.3, this research concentrates on the analysis of upper and lower regions of the face. The deterioration confidence rate corresponding to the percentage of “true deteriorated” responses was calculated for each condition of the five classes defined in the study of Madrigal-Garcia et al. The left avatar of Figure 4.3(a) displays a neutral expression, while the right avatar reveals a final-stage deterioration condition (AUs recorded at 100% of their maximal contraction). As illustrated in Figure 4.3, deterioration intensity increased significantly with the amplitude of certain action units (AU15, AU25, AU43, AU55, and AU56).

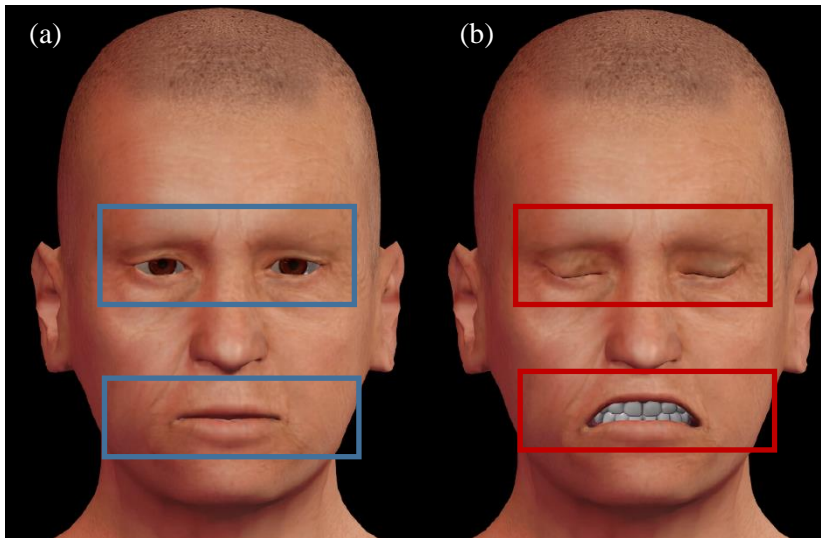


Figure 4.3 Facial expression areas that reveal if the patient is under deterioration or not.

(a) The left avatar expresses a neutral expression.

(b) The right avatar reveals deterioration status in the final stage.

Figure 4.4 illustrates the AUs of deterioration as defined by Madrigal-Garcia. In the picture, combinations of AUs define various face displays (FD). These FDs are the focus of this thesis, as certain combinations of FD provide a clear means to objectively quantify trend analysis.

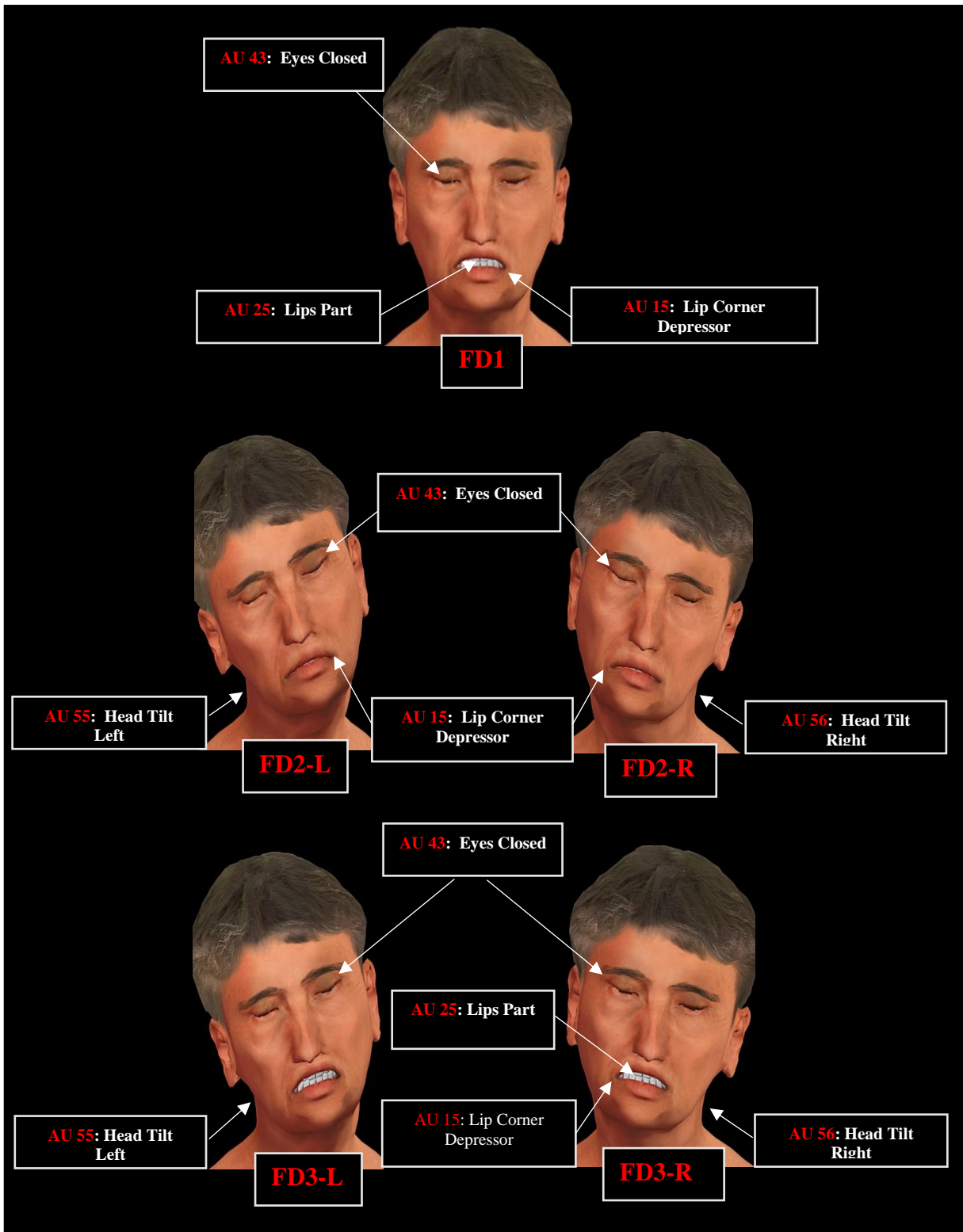


Figure 4.4 Expressive images and their active AU coding based on FACS. The composition of describing five different facial expressions represent patient at risk of deterioration using a collection of FACS based descriptors.

4.3 Generating the Dataset

The first and the most challenging task for FER is generating the dataset, which is considered the bottleneck in the process of designing, implementing, and developing deep learning models. The proposed FEs dataset was generated using a virtual human, a computer-generated 3D digital representation that looks and acts like a real human (Bombari et al., 2015).

The following sections will highlight the methods employed to generate synthetic data-driven FEs named Patient at Risk of Deterioration-Facial Expressions (PRD-FE).

4.3.1 Generating a Facial Expression Dataset via Avatars

Because there is no availability of real FER samples representing patients at risk of deterioration that include environmental illumination changes, this thesis presents a new benchmark generated dataset called PRD-FE, Patient at Risk of Deterioration Facial Expressions. Various methods are exploited to present an as realistic as possible dataset that represents FEs of patients in deterioration. Important factors have been addressed, such as producing a balanced dataset from which AI models can learn discriminative features under various conditions.

Research on creating FEs through avatars using computer-generated animated characters has significantly increased over the last few years and is now considered a valuable tool in studying emotions and social cognition (Treal et al., 2020). Using avatars allows highly controllable, interactive experiments, can provide diverse facial expressions, saves cost and time compared with human experiments, and encompasses a variety of data including ages, skin tone, and ethnicity. However, the limitations and drawbacks of avatars have to be considered. They cannot mimic or capture the richness and complexity of real facial expressions, especially the fine and subtle facial movements of micro-expressions. These shortcomings minimise the realism of avatars.

There are some recent studies that use 3D computer-generated to generate FEs. The work of (Buisine et al., 2014) showed an avatar expressing emotions in three conditions: still, idle, and congruent. Negative emotions in avatars were judged to be more intense when the character adopted a posture in congruence with its FE compared to a condition where the avatar was still.



The study of (Treal et al., 2020) explored the interaction between body motion and FE in term of intensity and believability of an avatar's pain expression. The work of (Sollfrank et al.,




2021) investigated whether the cortical response to emotional facial stimuli is influenced by the digitization of facial stimuli, referred to as "avatars," in a cohort of 25 individuals. Participants were presented with 128 static and dynamic FEs portrayed by both male and female actors, and their corresponding avatars, under neutral or fearful conditions. Various other studies have generated FEs through 3D avatars, including (Mukashev et al., 2021) who developed two methods. The first exploited tuning of Blendshape features of the 3D model for facial generation, and the second captured six basic FEs from a real face which was mapped onto the 3D model. Additionally, they used lip synchronization software to generate realistic lip movements.

In this thesis, the generation of dynamic FEs emulating the risk of deterioration is based on FACS (Ekman & Friesen, 1978). The FE dataset was generated according to the method of (Madrigal-Garcia et al., 2018). Here, we addressed automatic recognition of AU sets referred to as face displays (FD), as illustrated in Table 4.1, by adapting ML and DNN algorithms to recognise FEs in the high-dimensional space of the FEs through three-dimensional avatars.

The AUs of five FE classes, the muscles responsible for their occurrence, and image samples of these expressions are shown in Table 4.1.

Table 4.1 Action Units of five expressions at risk of deterioration and their relevant facial muscles.

Action Unit	FACS Name	Facial Muscle	Example Image
15	Lip Corner Depressor	Depressor anguli oris (Triangularis)	
25	Lips part	Depressor Labii, Relaxation of Mentalis (AU17), Orbicularis Oris	

43	Eyes Closed	Relaxation of Levator Palpebrae Superioris	
55	Head Tilt Left		
56	Head Tilt Right		

Generating FEs through avatars involves several stages, including the creation of AUs and the synthesis of FEs. Twenty-five avatars aged between 18 and 70 (average 30.05 years), 15 women representing 60% of the samples (M_{age} 41.2, SD_{age} 18.8368) and 10 men 40% of the samples (M_{age} 44.5, SD_{age} 17.8395) were generated with various skin tone (white, black, yellow), ethnicity (African, Asian, White), face shape (oval, long, square, heart, diamond), and facial features (eye and hair colour, shape and size of the nose, chin, cheeks, eyes, forehead, and lips). Each avatar implements five different expressions (FD1, FD2-L, FD2-R, FD3-L, FD3-R) labelled into five classes representing the patients at risk of deterioration.

The generated dataset consists of 125 video clips covering the 5 FEs and each video lasts around 11-12 seconds displaying animated faces at 25 frames per second. The avatars face the camera at various angles showing dynamic FEs representing risk of deterioration. There is no body movement in the video sequences, only head motions. The raw dataset has around 37,000 frames for the videos of all classes.

Because the avatar-generated FEs are to be recognised by automatic algorithms, the dataset contains images and videos recorded with a neutral background and normal lighting conditions. Such conditions can be emulated in the real world using a face detection algorithm and ensuring that the camera sensor is positioned correctly with acceptable illumination. In hospital wards, there may not always be enough lighting due to patient preference or comfort. One of the best solutions for this issue is using cameras with an infrared factor as IR light is invisible to the eye. Cameras have near-IR sensors and lighting, so they can flood the area with near-IR light if illumination is below a certain threshold.

4.3.2 Generating Data-Driven Facial Expressions using FACS

FACS was developed by Paul Ekman and is primarily employed in studies related to nonverbal human communication and activity and in the development of avatars and artificial agents (artificial facial animation) (Bennett & Šabanović, 2014; David et al., 2014). The FACS system breaks down FEs into individual component AUs, which correspond to specific movement of facial muscles (refer to Figure 4.5).

FACSHuman is a software tool based on FACS which, in conjunction with MakeHuman software, helps to produce experimental content like images and animations with high standards of realism, aesthetics, and morphological precision. This software enables configuration of the shape of face and body, and manipulation of the facial muscles, eyes, and tongue, etc. In addition, facial features, skin tone and eye colour, and the skeletal structure can be customised with MakeHuman software (Gilbert et al., 2021).

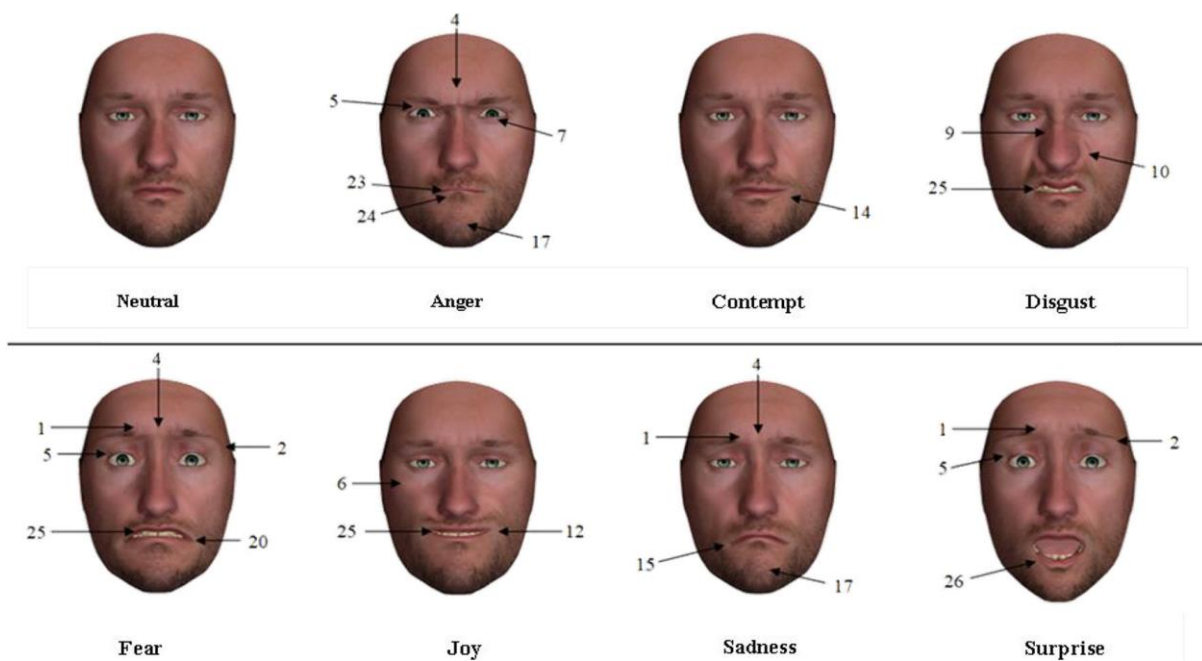


Figure 4.5 Sample of coded AU with FACS (Gilbert et al., 2021).

MakeHuman is a free and open-source software program based on open GL. It is used for generating realistic 3D human avatars as it has a 3D rendering engine. It is widely employed in the creation of 3D avatars in video games and 3D recreation tasks (Gilbert et al., 2021). The anatomy of avatars can be sculpted and manipulated, adjusting parameters such as height, weight, and facial landmarks.

Furthermore, MakeHuman offers parameters and tools to model the entire human body, including the face (shape of face and landmarks, eye colour, etc.), gender, skin tone, ethnicity, and age from infant to elderly, offering unlimited combination possibilities. The avatar identities can be created, saved, shared, and reused in various experiments. In addition, the generated AI agents can be exported into other 3D software such as Blender or Maya to address further objectives based on the task target. Unlike previous software tools limited to facial or upper body modelling, MakeHuman supports modelling in terms of body shape and posture (Gilbert et al., 2021).

All objectives addressed in FACSHuman plugins were generated using Blender software and subsequently imported into MakeHuman. These plugins make diverse objectives accessible and effortlessly adjustable. MakeHuman presents three important supplementary plugins developed using Python to provide advanced privileges to this software and a significant flexibility for modifying the source code or incorporating additional functionality as required. The first one is for crafting facial expression, the second one for manipulating facial features, and the final one for scene editing. These plugins produce the possibilities of creating images of various resolutions and degrees of intensity based on task target, to produce compile of static images to generate videos.

The first plugin is for crafting facial expression, the second for manipulating facial features, and the third for scene editing. These plugins facilitate the creation of images of various resolutions and degrees of intensity to compile static images to generate videos. Intricate FEs can be produced, and the movements of facial muscles, skin, eyes, jaw, and head can be defined, as shown in Figure 4.6.

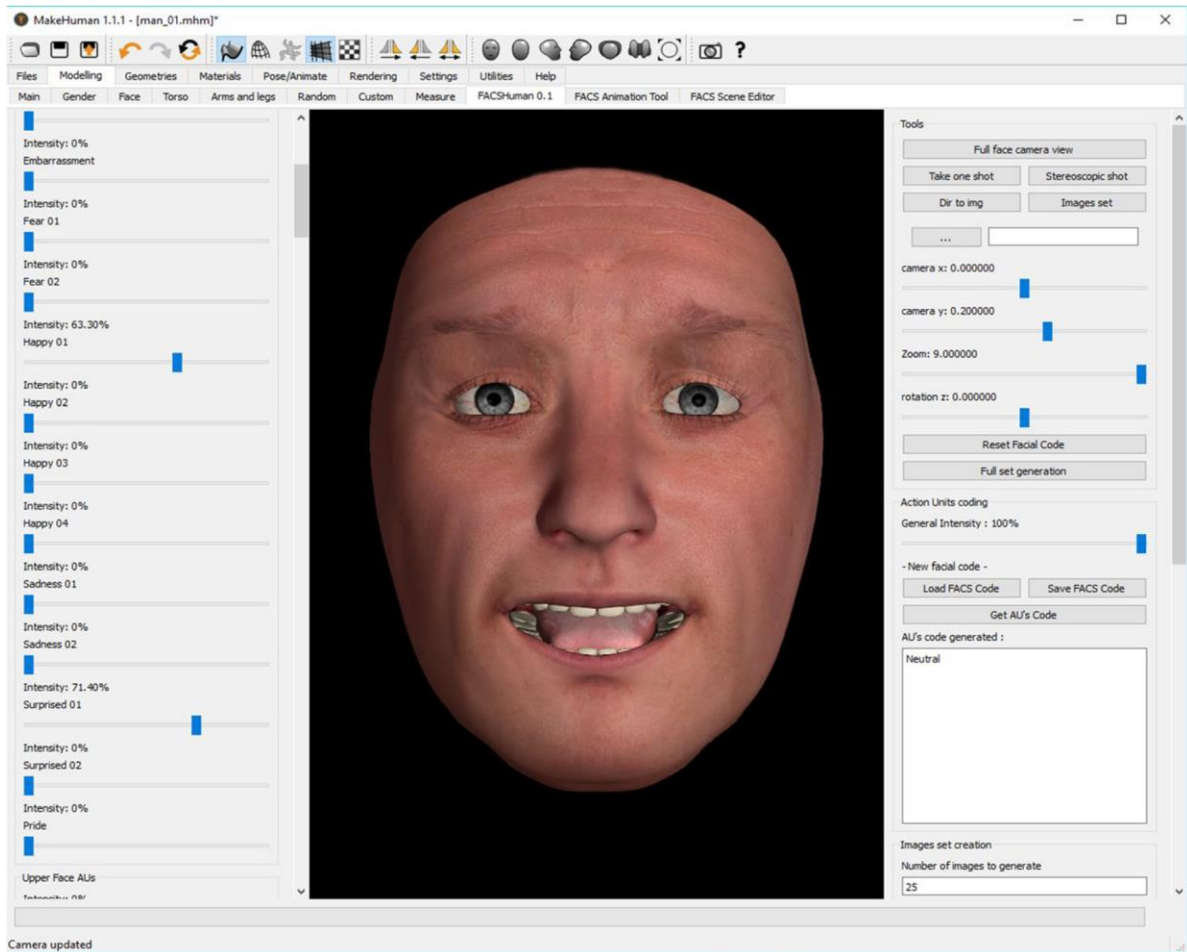


Figure 4.6 FACSHuman user interface (Gilbert et al., 2021).

The interface app offers tools for manipulating predefined FEs by adjusting the intensity of each AU using sliders. The app is also provided with tools to control the camera angle, zoom function, and position, and to load and save the FEs. Furthermore, it offers the possibility to program parameters for batch picture processing by specifying the number of produced images. This feature enables the gradual variation of FE intensity, useful for experiments involving recognition thresholds as shown in Figure 4.7.



Figure 4.7 Progress of facial motion intensity (Gilbert et al., 2021).

The avatar can combine and animate various AUs over a time frame to generate macro and micro FEs, as well as interactive animated FEs. Moreover, the produced videos and images have a transparent background, enabling modelled faces to be presented with coloured backgrounds or images. In addition to manipulating AUs, the plugin includes an emotion mixer, allowing the blending of various FEs, as highlighted in the study by Ekman (Ekman & Friesen, 1978). This plugin provides access to a wide range of configurations to control various AUs that are explored in the study of FACS.

FACSAnimation Tool (plugin 2) is designed for producing animated FEs, creating, configuring, and recording animated FEs either by generating each AU or by blending expressions crafted in the FE creation tool. The FE intensity can be adjusted based on a predefined expression. The duration of the video and number of images can be selected during batch creation processing, and with higher numbers of frames in the video animation, the video playback is proportionally slowed down due to the frame rate.

The FANT plugin features a parameterizable timeline, enabling the generation of intricate expressions. The AU intensity is generally categorised into three temporal regions (Ekman et al., 2002): Initial (onset), apex, and offset. The initial region represents the start of muscle contraction, the apex region refers to the plateau of maximum intensity during their progression, and the offset region describes the relaxation of the facial muscles. The flexibility to define the start and stop intensities for apex regions allows the generation of different intricate movements, as depicted in Figure 4.8.

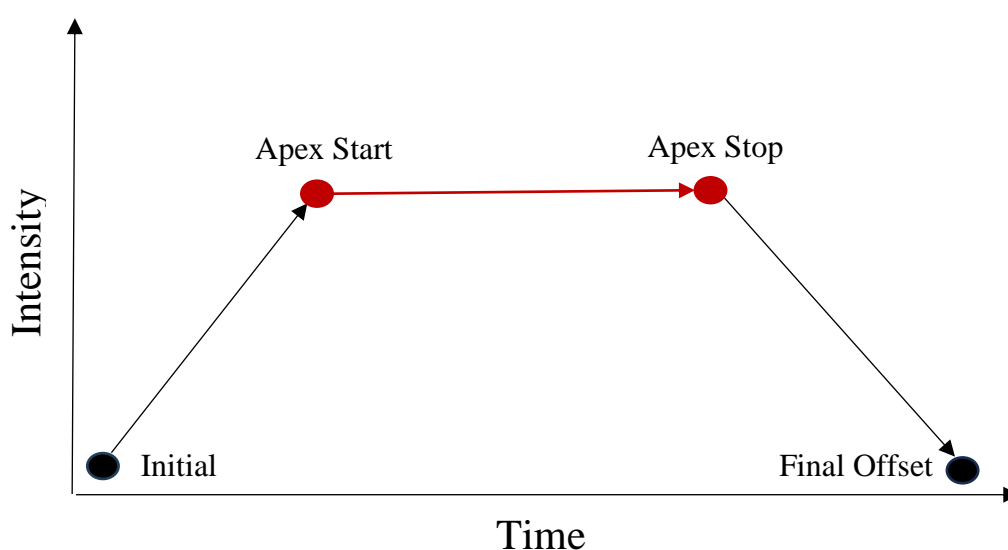


Figure 4.8 Attributes of an AU indicated on the timeline (Gilbert et al., 2021).

In addition, this plugin offers various tools to produce non-linear complex expressions which closely resemble human facial movements. Various AU attributes can be controlled, such as duration, start time, end time, and intensity and based on the timeline, diverse FEs can be produced.

The FANT plugin presents various AUs that can be exploited based on the target FE, offering timeline attributes such as start, end, and duration. Figure 4.9 depicts the FANT plugin facilities that produce expressions through controlling and manipulating different AUs.

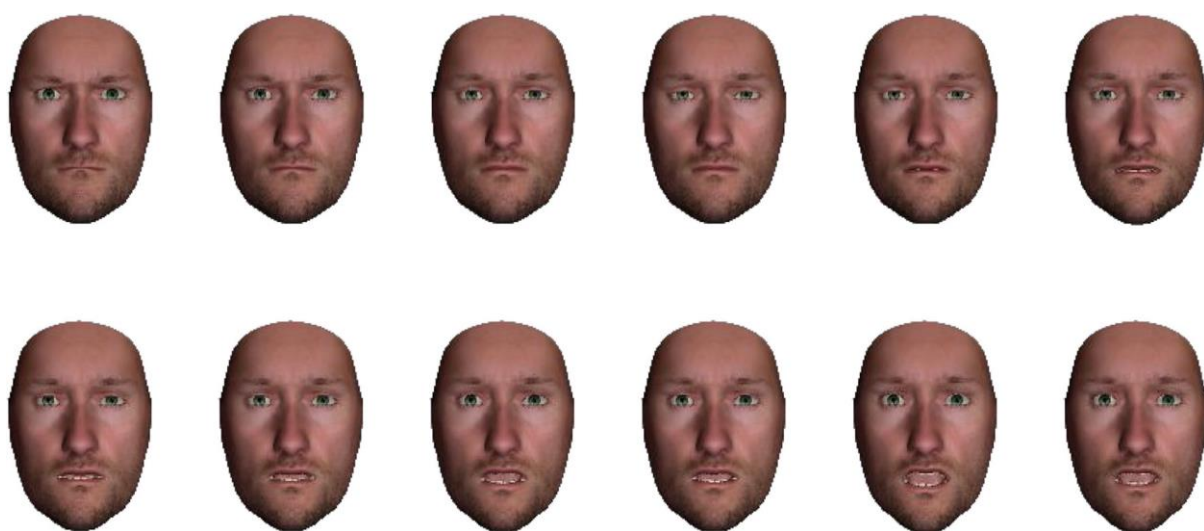


Figure 4.9 The sequence of images, generated using the FANT plugin, shows a transition from anger expression to surprise expression (Gilbert et al., 2021).

To produce realistic FEs that closely resemble human expressions, this software provides various AU features through these plugins such as degree and evolution of intensity, start and end point, and duration of movement. By controlling these AUs attributes, realistic synthetic animated FEs can be produced that imitate real samples with observable asynchronous movements.

AU intensity typically progresses through several stages, reflecting the degree of muscle activation in the face. These stages can be described using labels that range from minimal to maximal intensity. The typical stages of AU intensity progression are the following.

1. Initial (neutral). No activation with the face in a relaxed state.
2. Trace. Minimal activation reflected by slight and barely noticeable facial movement.
3. Slight. Light activation by slight facial movement that is more noticeable than trace.
4. Pronounced. Clear and distinct facial muscle movement with a noticeable appearance of FE.
5. Extreme. Significant alteration of the FE by strong muscle movement.
6. Maximum. Peak of intensity, expressing the maximum appearance of the FE.

Figure 4.10 presents an example of random AUs with different start and end positions and with varying levels of intensity.

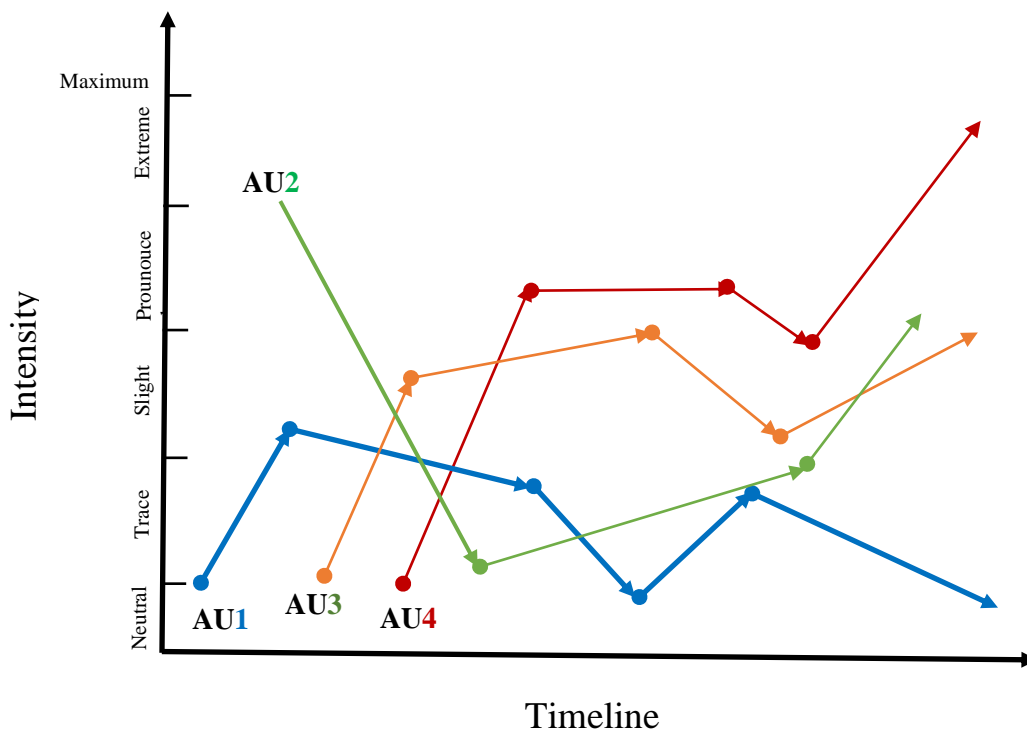


Figure 4.10 An analysis of the chronological sequence and layout of AUs intensity. Example of different AUs starts and ends at different time along with various intensity.

The progress of AU intensity based on a timeline has more influence on the perception of FE progress than total duration time (Kamachi et al., 2001), so the ability of ANN models to identify AU stage on a timeline is important for FER.

AU onset and offset duration were set to 4000 ms (at 25 fps), and the apex duration of the FEs was set to 7000-8000 ms, which includes 175-200 frames at 25 fps.

Figure 4.11 shows the progress of the synchronised AUs based on timeline. The five rows illustrate the progress of the five sets of AUs that represent the five types of FE of deteriorated patients, numbered 1 to 5. During onset, corresponding to the trace stage of AU progress, the five FEs start with minimal activation reflected by slight and barely noticeable facial movement. After 4000 ms, the AUs are lightly activated by slight facial movements that are slightly noticeable. Between 4000 and 8000 ms, the movement of facial muscles becomes clearer and more distinct with a noticeable appearance in FE until reaching extreme intensity resulting in a significantly altered FE. The final stage after 12000 ms, the AU intensity reaches its apex expressing the maximum appearance of the 5 FEs.

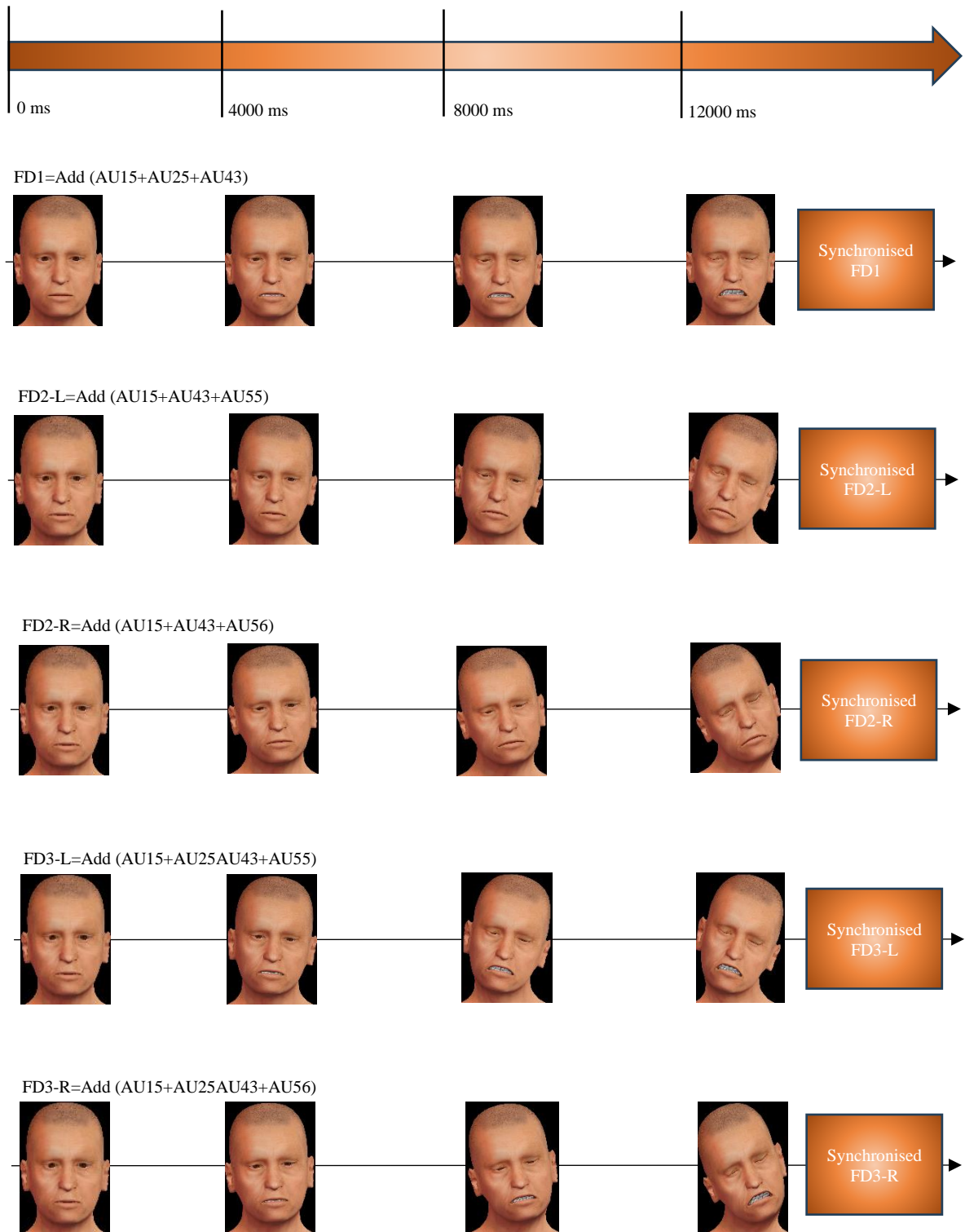


Figure 4.11 Description of progress of the synchronized AUs based on timeline.

The FACSSceneEditor plugin is responsible for configuring the lighting of the scene. Like a professional photographic studio setup, it determines lighting intensity, precisely specifying positions of light sources around the face through the x, y, and z axes. The selection of brightness, contrast, colour, specular reflection, and other characteristics supported by OpenGL allows for the creation of various staging effects and adjustments of image properties (Figure 4.12). These attributes ensure consistent lighting within the scene and a stable environment for image production.



Figure 4.12 Scene editor and lighting possibilities.

Batch creation of still images enables the production of videos with customizable frames per second and a pause interval. It also offers the advantages of specifying starting time, adjusting AU intensity, and generating a batch of images starting from the specified time.

This software and its plugins are an open source that can be used and updated under the conditions and agreements outlined in the license. Furthermore, this software and its plugins are compatible with most operating systems, ensuring accessibility and usability across various platforms. Table 4.2 shows the 5 classes of AU combinations with their descriptions, and several generated videos in each class.

Table 4.2 Combination of AUs of each class to form facial expressions of participants at risk of deterioration & Number of generated videos.

Expressions	Involved Action units	Description	Samples of video clips
FD1	AU (15+25+43)	Lip Corner Depressor, Lips part, Eyes Closed	25
FD2-L	AU (15+43+55)	Lip Corner Depressor, Eyes Closed, Head Tilt Left	25
FD2-R	AU (15+43+56)	Lip Corner Depressor, Eyes Closed, Head Tilt Right	25
FD3-L	AU (15+25+43+55)	Lip Corner Depressor, Lips part, Eyes Closed, Head Tilt Left	25
FD3-R	AU (15+25+43+56)	Lip Corner Depressor, Lips part, Eyes Closed, Head Tilt Right	25

The generated avatar videos simulate the behaviour of patients at risk of deterioration while the body is static, with or without the movement of head poses. The intensity of the deterioration expression was fixed at 100% of the maximal AU contraction depicted in Table 4.2. Each video began with the avatar showing a neutral expression (sets of AUs at 0). The AU intensities linearly increased to 8% contraction for 1 s.

The deterioration expression was maintained for 9 s until the end of the clip, as illustrated in Figure 4.13. In the first 4-5 s of each video, the avatar first stayed static with a neutral expression, then the facial muscle movement started to show a deterioration FE until reaching a maximum at the end of the video. The automatic model analyses each frame and identifies the frame in which the FE indicates deterioration risk. The position of the cursor along the scale was converted to numerical values between 0 (normal condition) and 100 (worst deterioration condition).

The avatar's deteriorated expression was perceived to be more intense and more believable in a combination of the upper and lower parts of the face. In Figure 4.12, an avatar assumes the five different expressions of deterioration, each with an AU set. The set AU15+25+43 describes the class named Face Display 1 (FD1), AU15+25+55 describes FD2-L, AU15+25+56 describes FD2-R, AU15+25+43+55 describes FD3-L, and AU15+25+43+56 describes FD3-R.

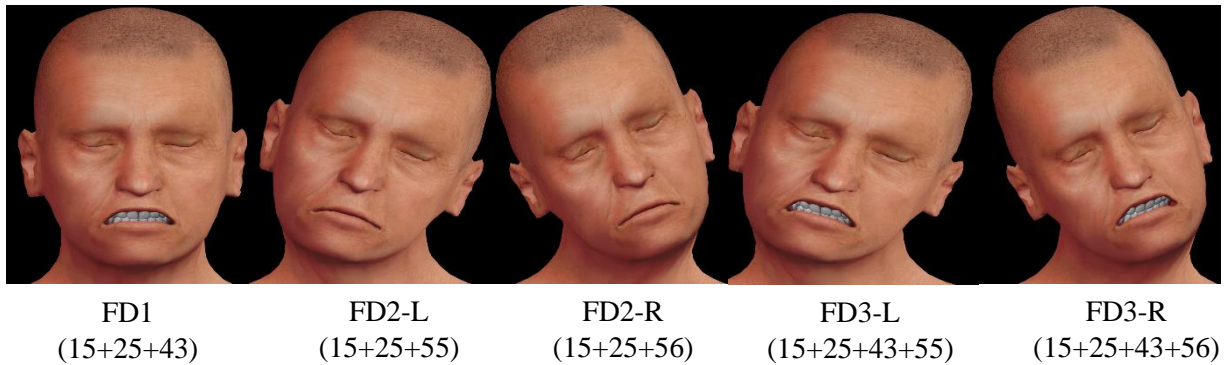


Figure 4.13 Five classes along with the combination of Action Units.

4.4 The Transfer of Facial Expressions to Static Real Faces using the First Order Motion Model (FOMM)

Face swap, bring the face to life, image animation, and Deepfake generation techniques are applications used to replace the face of one person with the face of another in a sequence. The First Order Motion Model (FOMM) is a DL framework designed by Siarohin et al. (Siarohin et al., 2020) for generating realistic motion in images. The framework generates video sequences in such a way that the object in the source image is animated according to the motion of the driving video. FOMM is typically used for tasks such as animating a static portrait or synthesising video frames from a single image. This model leverages the first-order motion of key points and their local affine transformations to produce high-quality animations. The ability of this model to learn FEs is significant, without the need for prior information about the object to animate. In image animation, a video is synthesized from two main parts. The first is the source image, the second the driving video. The model is trained on a dataset of images and videos of objects of the same category (e.g. face, body) by identifying key points on the object and then pegging them to the motion in the videos. Recent applications of CNN have realistically mimicked human faces, and training networks on many images and video datasets can generate realistic talking faces.

In the FOMM model, a source image of a person can be animated to the target poses of another in the driving video (Malik et al., 2020). FOMM combines the appearance extracted from the source image and the motion derived from the driving video. The framework described in

(Siarohin et al., 2020) has achieved satisfactory results on a variety of object categories. Their model has pre-processed the dataset, extracting an initial bounding box in the first video frame. Then it tracks the object until it is distant from the initial position. After that, video frames are selected with the smallest crop containing all the bounding boxes. This process is repeated until the end of the sequence. Sequences with a resolution lower than 256×256 are filtered out and the remaining are resized to 256×256 , preserving the aspect ratio for a more realistic video in which the head moves freely within the bounding box. The model uses 19,522 training videos and 525 test videos, with lengths varying from 64 to 1024 frames.

Recent studies have used FOMM for different objectives. For example, in 2021, the work of (Y. Zhang et al., 2021) introduced a model combining the facial-prior-based method and first order motion model for generating facial micro-expression. Given a target face, the model drives the face to generate micro-expression videos according to the motion patterns of source videos. This model consists of three stages. The first extracts facial prior features from a specific part, then the second estimates facial motion using key points and local affine transformations with a motion prediction module. Finally, the FE generation module is used to drive the target face to generate new videos. The model is conducted on public CASME II, SAMM, and SMIC databases, and is then used to generate new micro-expression videos for evaluation.

This project has adapted FOMM to capture FEs from various images. The model is trained to reconstruct training videos by combining a single frame with a learned potential characterisation of the motion in the driving video. For testing, we apply our model to pairs composed of the source image and each frame of the driving video to perform image animation of the source object. The model is trained and tested with different datasets containing various objects. The method automatically produces videos by combining the appearance extracted from a source image with motion patterns derived from a driving video. For instance, a facial image of a person can be animated following the facial expressions of another.

The method is employed in this thesis to generate a dataset of realistic faces through transferring involuntary FEs, and head poses to detect patients at risk of deterioration. FE, eyeball movement, and head pose of a real face in a source image can be animated based on the motion of an avatar FE and head pose in a driving video (Figure 4.14).

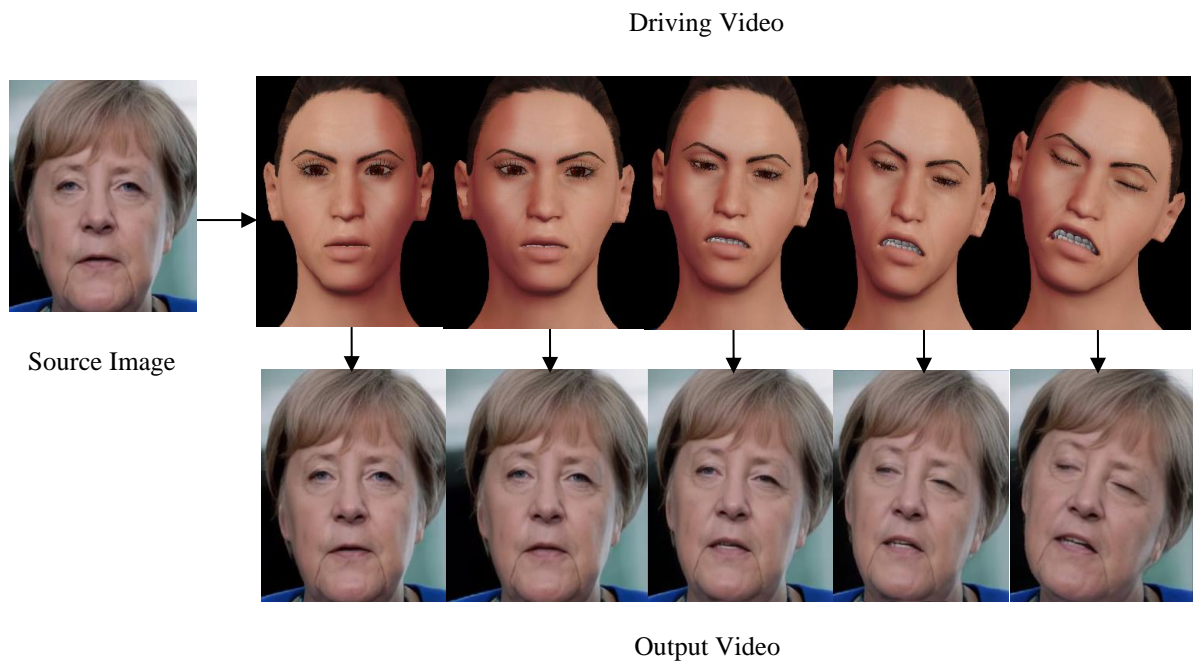


Figure 4.14 Frames of Video Sample After Utilizing FOMM to transfer facial expressions from avatars.

FOMM was applied using the driving video of a 3D animated character displaying the specific FEs of a patient in deterioration. These expressions were faithfully mimicked on the various source images of real people's faces, as shown in Figure 4.16. The model is implemented using source images of real human faces from an open database available on Kaggle known as Celebrity Face Image Dataset, and the FEs, head poses, eyeball movement, and other actions from the videos transferred to the source images are considered of good quality and realistic. Transferring the FEs data to static real faces to bring them to life is essential for training the model for FER of real human faces. Creating a realistic robust model that can be applied in the real world on real faces can be achieved with a model that can bring real facial images to life, as can be seen in Figure 4.14.

In summary, the avatar-generated FEs were transferred to real facial static images in the Celebrity Face Image Dataset. Using FOMM, 176 video clips (see Figure 4.15) have been generated with colourful frame sequences of 11-12 second duration at around 25 fps, giving 275-300 frames in each video.

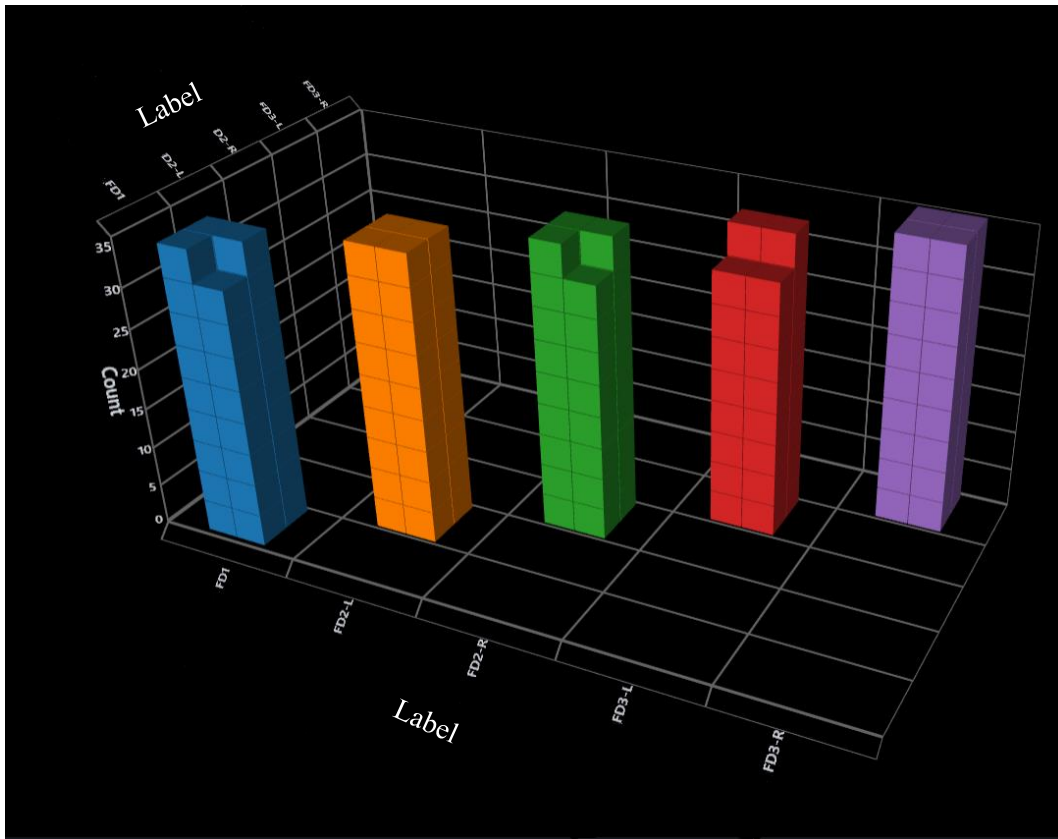


Figure 4.15 Bar chart depicts number of generated videos in each class of 5 classes.

Figure 4.15 shows a 3D bar chart with different colored bars (blue, orange, green, red, and purple) representing number of videos for five different categories labeled (FD1, FD2-L, FD2-R, FD3-L, and FD3-R) respectively.

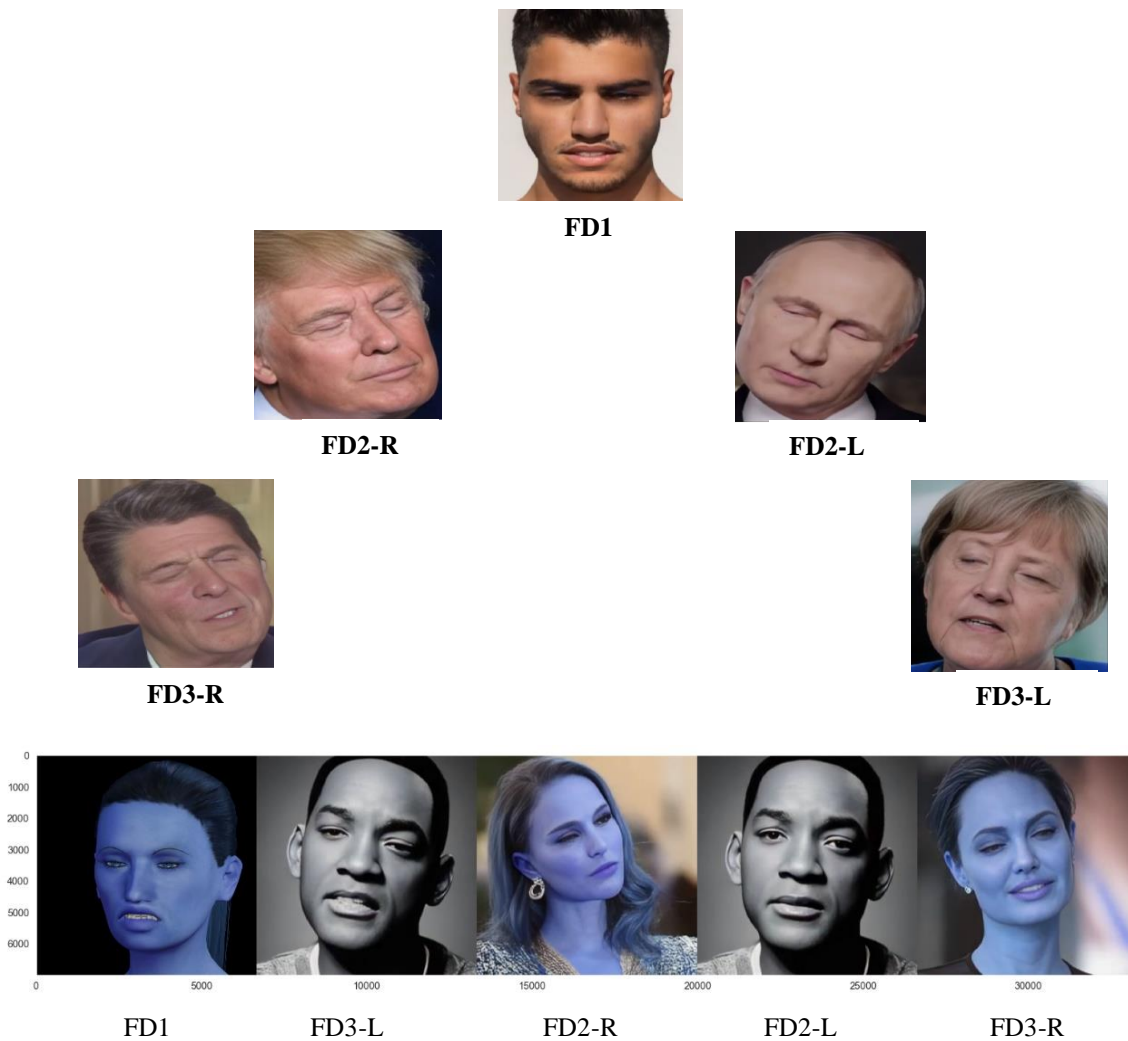


Figure 4.16 Samples of five classes.

4.5 Conclusion

Five FEs, including macro and micro expressions, have been generated through avatars based on work of (Madrigal-Garcia et al., 2018)

A synthetic database of diverse and realistic FEs has been generated using tools like FACSHuman software to create highly realistic 3D avatars capable of displaying FEs under a wide range of conditions, such as varying light, face scaling, camera angles, and backgrounds. This simulation of real-world conditions was crucial for producing a versatile and authentic dataset.

Key aspects of our methodology are the following.

1. Advanced modelling and animation techniques.

FACSHuman software was applied to craft 3D avatars. This involved detailed facial rigging and expression synthesis to mimic human facial expressions accurately. The flexibility of the software allows for the generation of FEs in limitless scenarios, enhancing the realism of synthetic data.

2. Comprehensive demographic representation.

A significant emphasis was placed on creating avatars that represent a diverse range of ages, genders, and ethnicities. This approach ensures that our dataset is inclusive and unbiased, providing a comprehensive tool for various applications in AI and ML, where diversity and representation are critical.

3. Realism through First Order Motion Model.

To bridge the gap between synthetic and real-world data, we used the First Order Motion Model. This technique enables the transfer of synthesised facial expressions onto static images of real human faces, producing a dataset that closely imitates real-world samples. This method not only enhances the realism of the synthetic data but also ensures that the generated FEs are lifelike and applicable in practical scenarios.

In conclusion, the combination of advanced 3D modelling, diverse demographic representation, and the innovative use of motion transfer models results in a synthetic database that is both realistic and representative. This dataset is poised to significantly impact fields such as facial recognition, human-computer interaction, and AI-driven animation, providing a robust foundation for future research and development.

Chapter 5

Facial Expression Classification Using Random Forest and Support Vector Machines

This chapter presents the thesis approach for modelling facial expressions recognition to enable automatic prediction patients' deterioration using traditional machine learning models, namely Random Forest and Support Vector Machines.

5.1 The Proposed Method

FER approaches are based on recognising and classifying combination sets of AUs within the input frames through AI algorithms. These algorithms can determine FEs of deteriorated medical conditions by interpreting the detected AUs. Figure 5.1 shows the flowchart of the methods used to design the proposed system. It starts with incorporating an infrared auto-tracking camera for face detection. The infrared capability is beneficial for improving detection in various light conditions, such as low-light or night-time detection. After detecting a face, the system verifies if the detection was successful. This could involve checking if the detected face meets certain criteria, such as size or orientation. Feature extraction methods are applied for extracting discriminative features that relate to facial landmark positions using the face mesh method and headtilt directions. This can be done using feature extraction methods like keypoint detection using convolutional neural networks (CNNs).

The final stage is feeding extracted discriminative features to RF or SVM classifier.

Based on the deterioration detection results, the system takes one of two actions: If the face components are stable and no deterioration is detected, the system continues monitoring. If deterioration is detected, an alert is created to notify healthcare professionals.

In summary, the proposed method provides a comprehensive approach to detecting and tracking facial components using infrared imaging, feature extraction, and traditional machine learning approaches. Detailed component detection and tracking allow for a nuanced analysis of facial expressions.

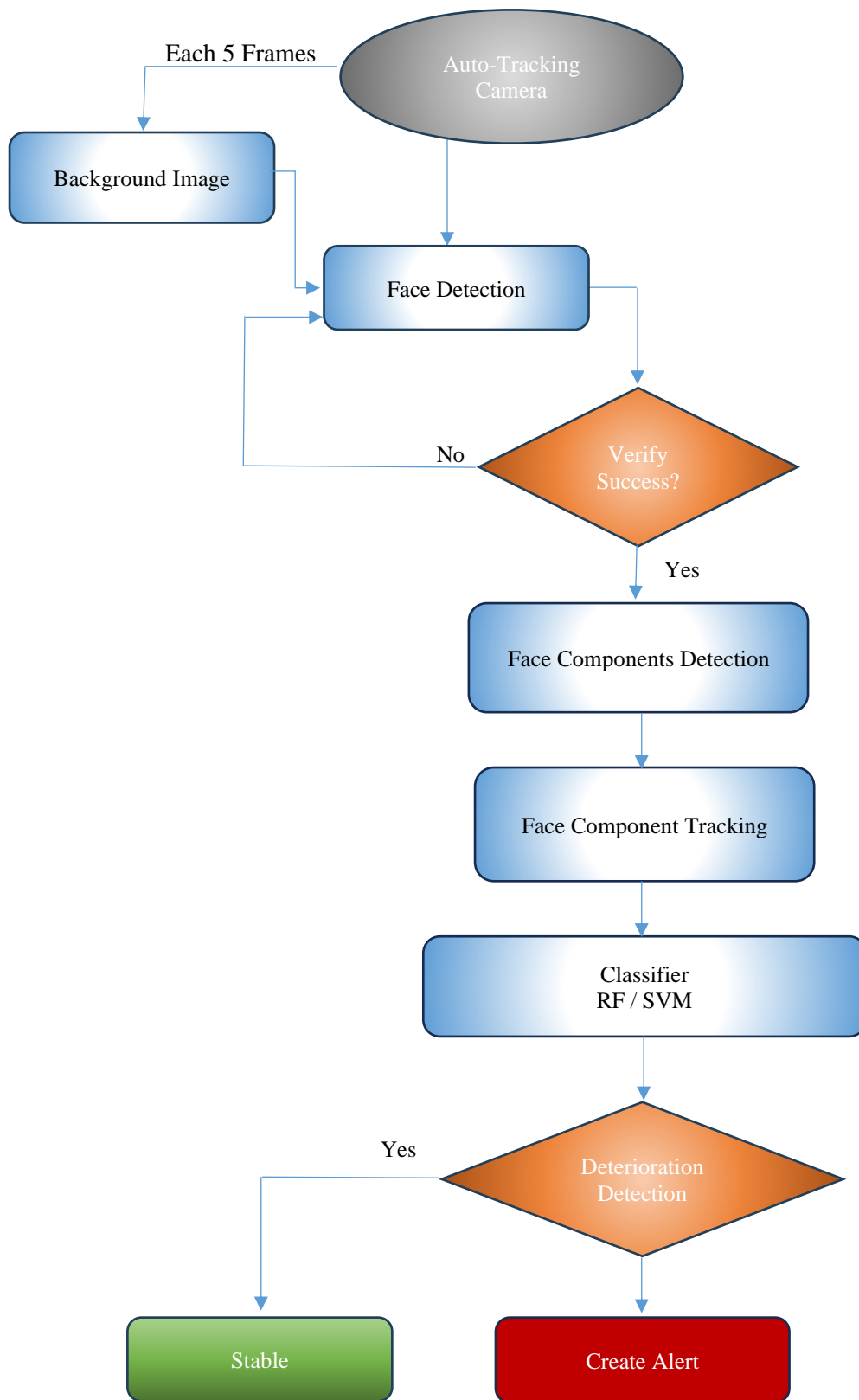


Figure 5.1 The proposed system flowchart based on facial landmarks as feature extraction method along with RF or SVM as classifiers.

The dataset is generated and expanded using avatar faces or 3D faces models that imitate the facial movement patterns of patients in stable and deteriorating condition. These FE movements are transferred to images of real faces using the deepfake model to make the system more realistic. Then the generated video data is preprocessed, extracting discriminative features as output data for categorisation. Finally, a meaningful comparison is carried out between the various algorithms applied to achieve a fair, acceptable, and accurate result.

Figure 5.1 shows the flowchart for the proposed methodology. Figure 5.2 illustrates the method making use of facial landmarks for feature extraction, the output of which is then used in connection with Random Forest or Support Random Machine as classifiers.

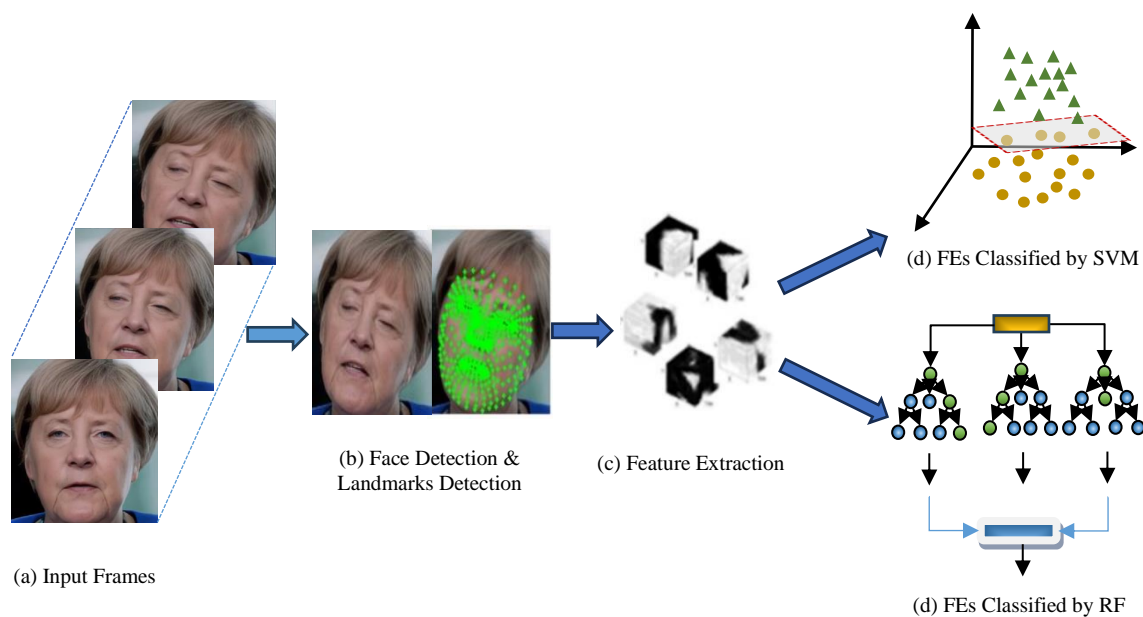


Figure 5.2 Main Stages of AFER based on feature extraction and Machine Learning. (a) input images. (b) face region and facial landmarks are detected, (c) Geometric features are extracted from the face components and landmarks. (d) the facial expression categories using either RF or SVM as Classifiers (Face images are taken from our generated database called PRD-FE).

5.2 Data Pre-Processing and Feature Extraction

5.2.1 Pre-Processing Techniques

In recent years, research on deep neural network-based FER has shown that these approaches overcome the limitations of conventional ML-based FER. However, as DNN-based FER approaches require an excessive amount of memory and incur high processing costs, their application depends on the hardware specifications and is very limited (Jeong & Ko, 2018).

In the preprocessing stage, facial images are resized to a fixed dimension with aspect ratio 4:5 for uniform analysis. After that, OpenCV libraries are used to convert images from BGR (Blue, Green, Red) to RGB colour space. This conversion is usually applied when processing images in OpenCV, which reads images in BGR format by default, especially when displaying images using libraries and tools like Matplotlib and PIL (Python Imaging Library) which expect images to be in RGB format.

In the next step, MediaPipe libraries are used to detect 468 3D facial landmark points in real time and to find distance-based features for categorising deteriorated FEs. The MediaPipe Face Mesh algorithm is a real-time facial landmark detection algorithm developed by Google's MediaPipe team. It is designed to accurately detect and track facial landmarks, key points on the face like the eyes, nose, mouth, and contours, in images and video streams. The algorithm uses a DNN-based approach for facial landmark detection. It first detects the face in the input image or video frame using a face detection model. Once the face region is localised, the Face Mesh algorithm predicts the coordinates of a set of predefined facial landmarks within that region.

A transfer learning approach can be used to generate 3D facial landmarks from images and videos by training a neural network. This approach results in a reasonable 3D landmark prediction for synthetic and real-world data. We can then exploit the features of the selected MediaPipe face mesh landmarks. Each landmark point is described by its (x,y,z) coordinate values.

Such a large number of geometrically described facial landmarks can capture the finest details to recognise the micro and macro FEs that DL models find difficult to distinguish. Then, distance-based feature vectors are calculated from the facial landmark coordinates. These features have been ranked using the binary combination method to find the most relevant. Finally, the ranked features are used for the classification of FE using the Random Forest classifier.

This thesis makes extensive use of the MediaPipe libraries, so further details are provided as follows. The MediaPipe Face Mesh algorithm is part of the MediaPipe framework, which provides a set of pre-built, customisable ML models and pipelines for various computer vision and multimedia tasks. It is designed to be efficient, accurate, and suitable for real-time applications on a variety of platforms, including mobile devices and desktop computers. There are several key steps involved in the MediaPipe Face Mesh algorithm.

First, the algorithm reads each frame from a video stream as its input. Then, it starts by using a pre-trained face detection model to align the region of face within the frame, which helps localise the region of interest for facial landmark detection. Once the face region is detected, the algorithm applies a DNN model to predict the coordinates of a set of facial landmarks (468 key points) within the detected face region. These facial landmarks typically include points corresponding to the eyes, eyebrows, nose, mouth, and facial contours. Because the proposed system works with video stream in real-time, the facial tracking detection exploits temporal information to track the movement of facial landmarks across consecutive frames. This helps maintain consistency and smoothness of facial landmark tracking over time. In a final step, landmark refinement may be applied to refine the detected landmarks by smoothing or filtering to improve accuracy and stability, especially in real-time applications.

Overall, the MediaPipe Face Mesh algorithm provides a robust and versatile solution for facial landmark detection, enabling a wide range of applications in areas such as computer graphics, human-computer interaction, and facial analysis, as illustrated in Figure 5.3.

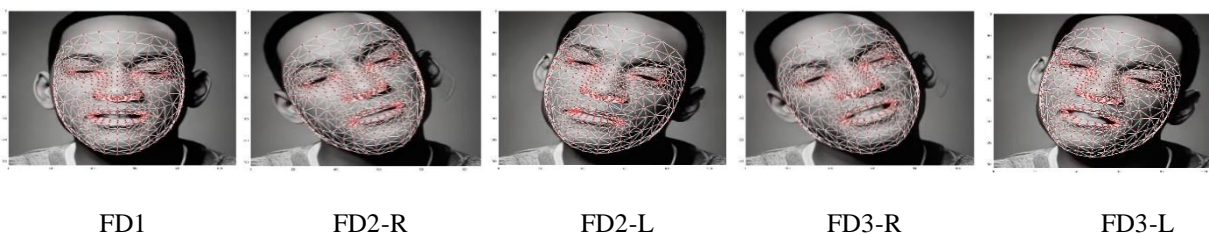


Figure 5.3 Facial frames samples for each class after pre-processed using a face mesh as face detection technique.

5.2.2 Feature Extraction

To recognise FEs in real time and minimise processing cost, the feature extraction method is based on the geometric features obtained through MediaPipe libraries. To detect head pose direction, for instance, the distance and the head tilt angle ratios between relative positions of key points are measured to extract features that represent side head tilting. To estimate head tilting direction, a function has been designed to extract the necessary key points to calculate relative distances and determine head pose location.

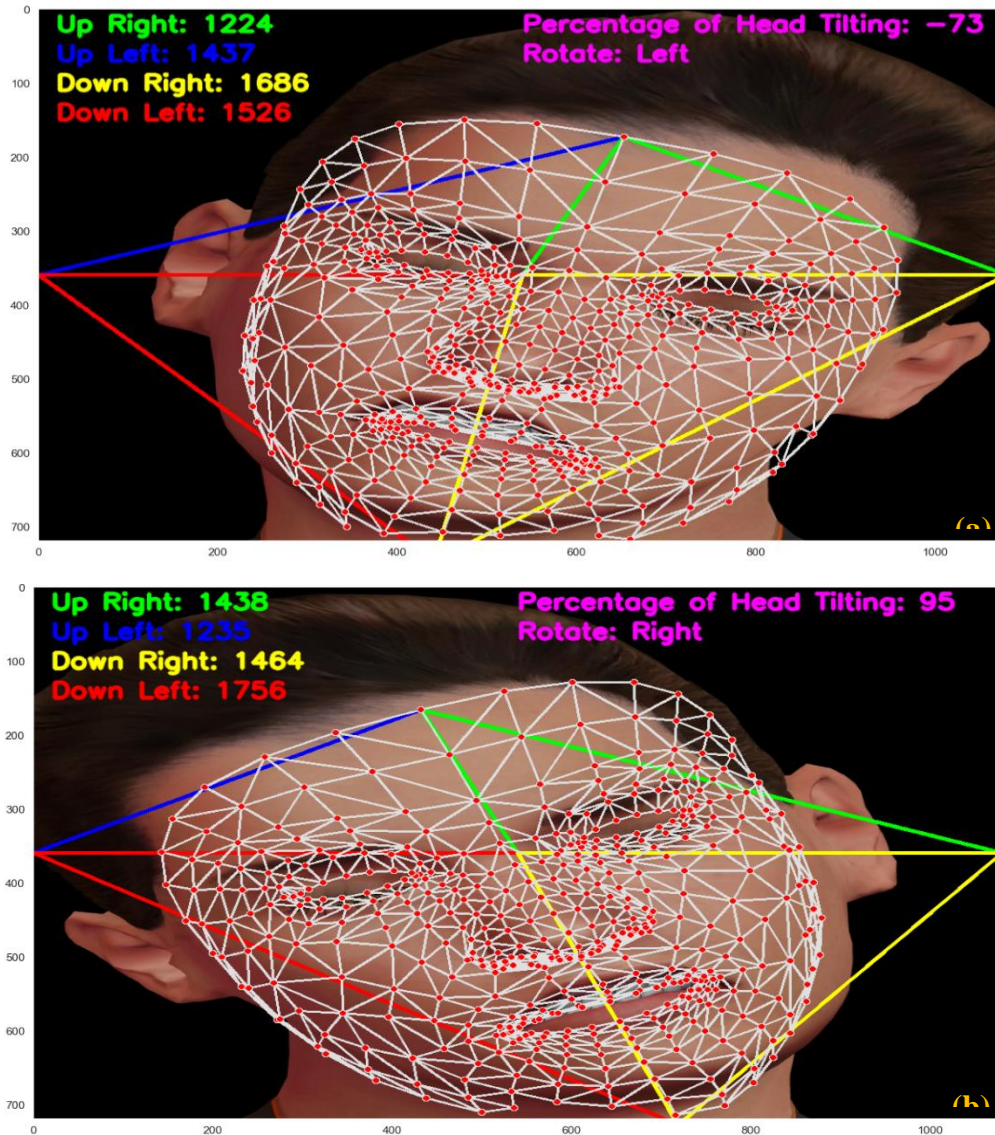


Figure 5.4 Four designed Triangles to extract features of head tilt direction.

As depicted in Figure 5.4, blue and red triangles mark the left side of the face, green and yellow the right. When the up-right and down-left triangles increase compared with the others, the head tilt is in the right direction (Figure 5.4a), while if up-left and down-right are larger than the others, the heads tilt in left direction (Figure 5.4b). Extracted features that include facial key points and head tilt measurement will be fed into the Random Forest classifier. The complete architecture of the proposed methodology for FER of patients at risk of deterioration is depicted in Figure 5.5.

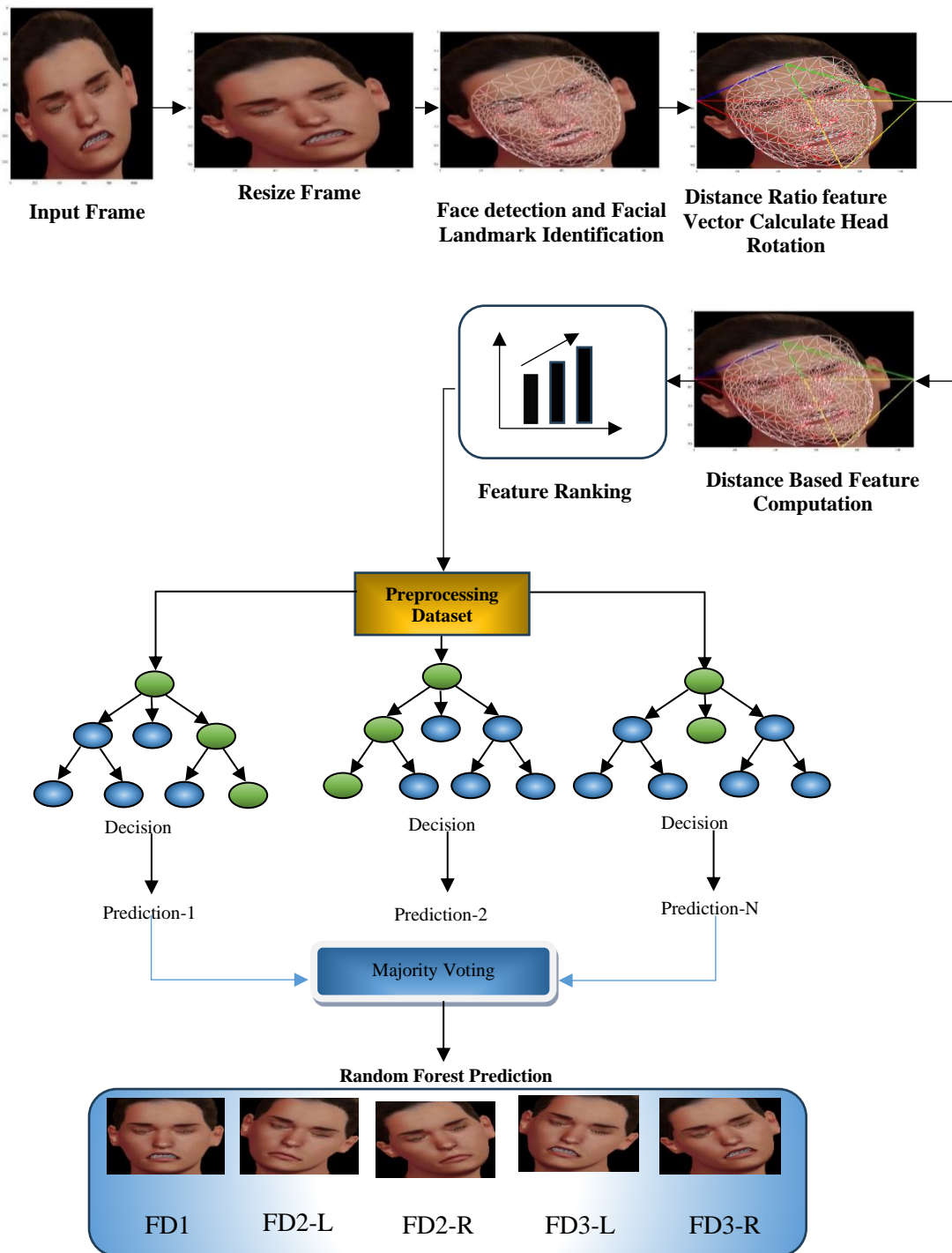


Figure 5.5 Overview of the proposed method for facial expression recognition. The images are resized, and face detected, then geometric features are extracted based on the distance ratio and angle relations for FEs and head tilting. Finally, the hierarchical weigh random forest classifies the facial expressions.

5.3 The Random Forest Classifier

Data was split into 85% training and 15% testing sets. The training stage starts with 2000 trees, a random state 42, and a maximum depth of 50 for each tree. The input features from an image or multi-dimensional format are reshaped (flattened) before being fed into the Random Forest (RF) classifier. The model is then trained using the flattened training data.

5.3.1 RF Model Results and Evaluation

The input data for the RF model are still images representing the various face displays (FD) represented in Figure 5.6. The data contained 195 samples for FD1, 187 for FD2-L, 183 for FD2-R, 185 for FD3-L, and 187 for FD3-R. Figure 5.6 presents the confusion matrix that summarises and visualises the performance of the proposed model. Each row of the matrix presents FEs in the actual class, while each column represents FEs in the predicted class. The proposed model attains 100% accuracy.

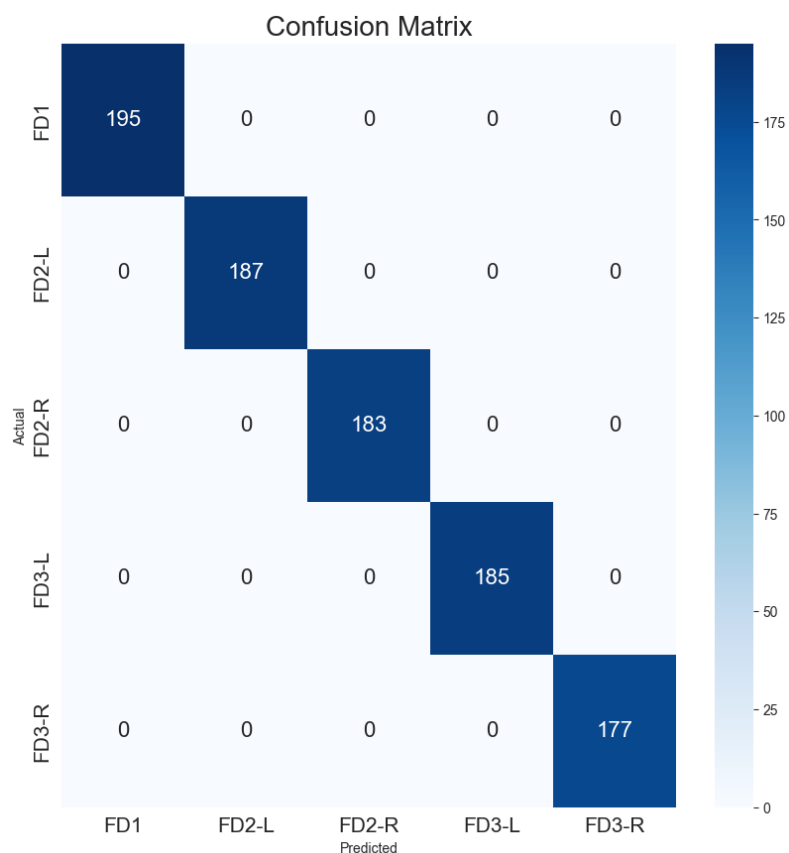


Figure 5.6 The Confusion Matrix of RF Classifier.

Figures 5.7 and 5.8 illustrate all evaluated measurements of the proposed model. Perfect classification is observed in terms of precision, recall, and F1 score.

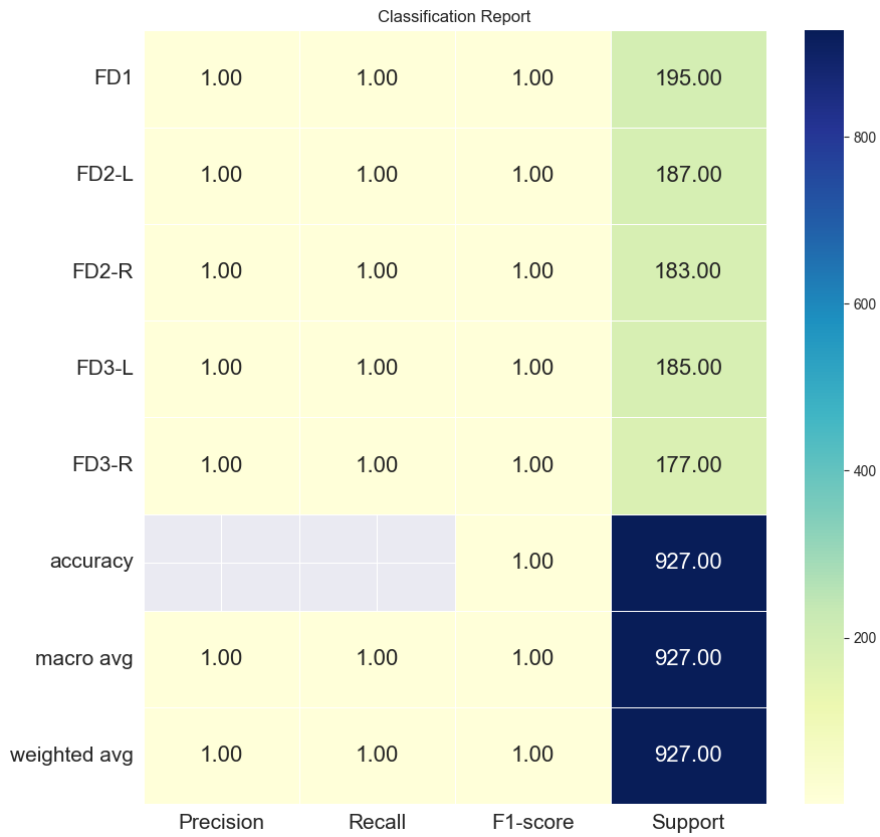


Figure 5.7 The Classification Report of RF Model.

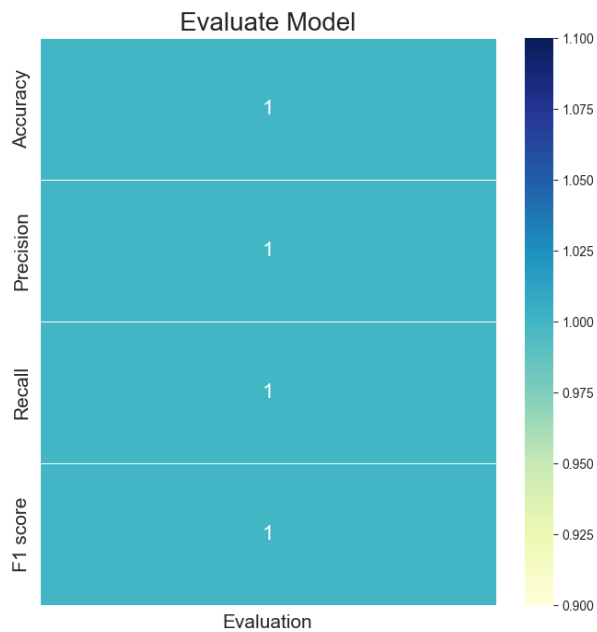


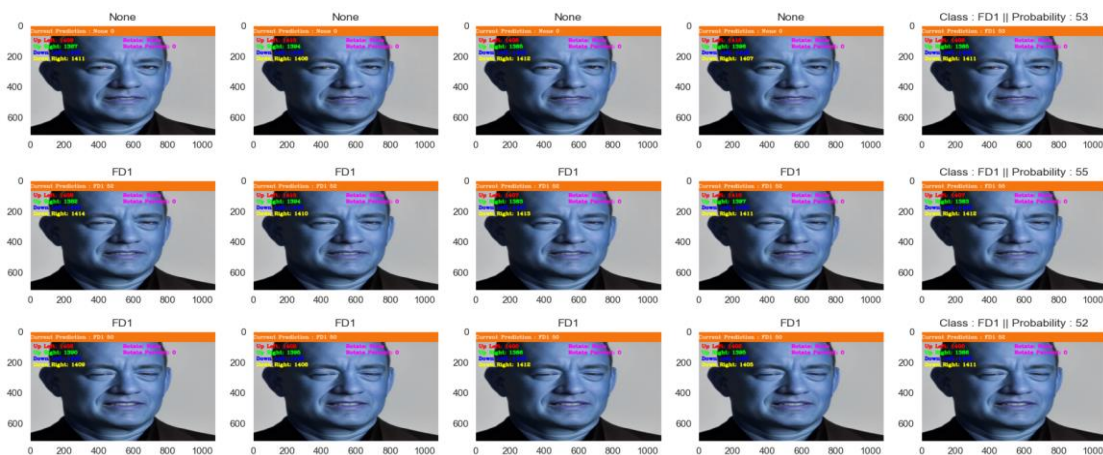
Figure 5.8 The Evaluation Metrics of RF Model.

The five evaluation measurements for five classes are illustrated in the following Table 5.1:

Table 5.1 Evaluation metrics for each Class & The Total Mean Performance of RF Model.

Class Name	Facial Expression	Precision	Recall	F1 Score	Accuracy
FD1	AU (15+43+25)	100%	100%	100%	100%
FD2-R	AU (15+43+55)	100%	100%	100%	100%
FD2-L	AU (15+43+56)	100%	100%	100%	100%
FD3-R	AU (15+25+43+55)	100%	100%	100%	100%
FD3-L	AU (15+25+43+56)	100%	100%	100%	100%
	Mean	100%	100%	100%	100%

The results clearly show that all FEs of interest can effectively be detected using face detection and face mesh of MediaPipe for facial feature extraction, and distance ratio for geometric feature extraction to determine head tilting direction, along with RF algorithm as a classifier. The above measurements provide insights into various aspects of model performance. The precision, recall, F1 score, and accuracy were all recorded 100%. In addition, the model has been evaluated using the unseen data separated from whole dataset before training, and the model also shows high predicted results as shown in Figure 5.9.



(a)



(b)



(c)



(d)



(e)

Figure 5.9 The RF classifier accurately classified unseen data into 5 categories.
 (a) Accuracy of prediction of unseen data predicted as Class FD1.
 (b) Accuracy of prediction unseen data predicted as (Class FD2-L).
 (c) Accuracy of prediction of unseen data predicted as (Class FD2-R).
 (d) Accuracy of prediction of unseen data predicted as (Class FD3-L).
 (e) Accuracy of prediction of unseen data predicted as Class (FD3-R).

5.3.2 Real-Time Model Evaluation

Figures 5.10 and 5.11 show real-time model prediction using the author's face with different FEs as unseen data which the model has not seen before (neither in training nor testing), and the model shows significant accuracy. However, there are still FEs in some facial frames that are not predicted because the author is not a professional actor, and due to the factors listed below.

Capture of transient expressions

A higher frame rate captures more frames per second, increasing the likelihood of recording fleeting expressions, such as micro expressions or quick emotional changes. These short-lived expressions might last only a fraction of a second but can provide significant insight into a person's true emotions. On the other hand, lower frame rates can result in missing brief expressions that occur between the frames. This can lead to incomplete or inaccurate interpretation of a subject's emotional state.

Temporal resolution and detail

A higher frame rate provides higher temporal resolution, offering more detail about how an expression evolves over time. This can be crucial for applications in which understanding the progression and subtlety of FEs is essential, such as in psychological studies or advanced HCI systems. Conversely, a lower frame rate reduces the temporal resolution. Important nuances and transitions in facial expressions may be lost, making it harder for FER systems to accurately analyse the dynamic movements of FEs.

Frame Rate and Motion Blur

At lower frame rates, fast-moving objects or rapid facial movements can appear blurred. Motion blur can obscure critical facial features that are important for accurate FE recognition, thus degrading the performance of FER systems. Higher frame rates generally reduce motion blur because the interval between each frame is shorter, so capturing movements can be more effective.

Computational Resources

As real-time processing demands immediate analysis and decision-making, limited computation resources may result in computational delays which may cause some frames to be skipped and not analysed.

In addition, latency, the time it takes to process and analyse each frame, can lead to delays causing subtle or quick changes in FE to be missed. Consequently, deploying models on edge devices that process data locally can reduce latency and improve real-time performance. Furthermore, frame rate may need to be balanced with the computational resources of the system to optimise the efficiency of real-time image tasks. Lower frame rates can ensure that the model can effectively perform frames analysis in real time without delay, but that can result in failing to capture brief micro facial expressions. On the other hand, a higher frame rate can capture sufficient FEs details but requires high specification devices with high computational processing ability and storage. In summary, choosing the right frame rate for FER involves considering the specific needs of the application, the nature of the FEs of interest, and the capabilities of the available computational resources.

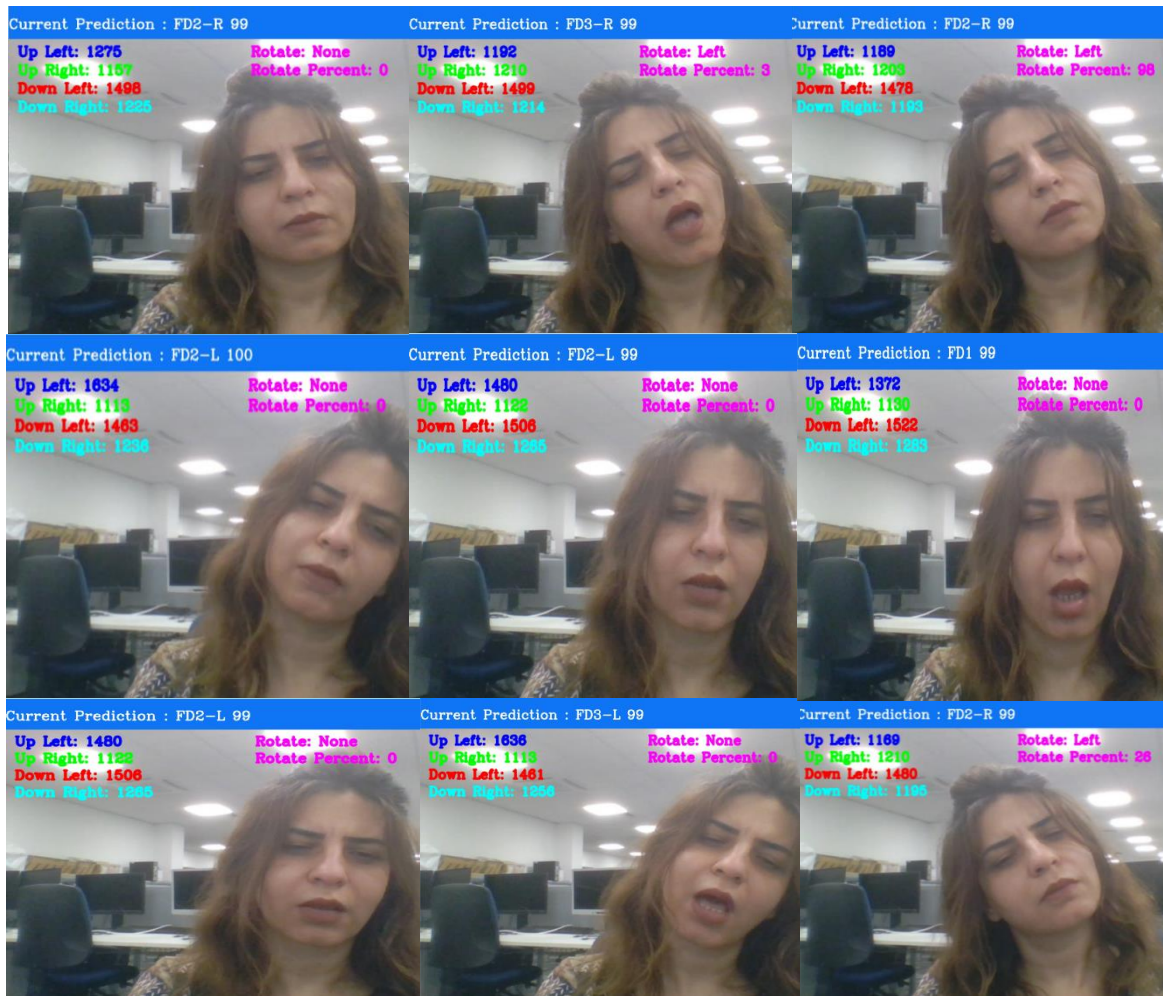


Figure 5.10 Shows real-time model performance on author face. The author tries to mimic FE of each class.

Improving the model's ability to consistently recognise expressions across all frames in real time involves addressing the limitations, the unpredictability of real-world conditions, and environmental challenges through both hardware upgrades and software optimisations.

Other FEs are also tested in real-time to test the model behaviour towards non-trainable expressions. If the model is properly trained with the dataset comprising the 5 FE types (FD1, FD2-L, FD2-R, FD3-L, FD3-R), it should not recognise any other FEs. Figure 5.11 shows other FEs expressed by the author in real-time, clearly demonstrating that the model could not recognise them. With prediction 0, the model can predict no FE that indicates the medical condition of the patient.

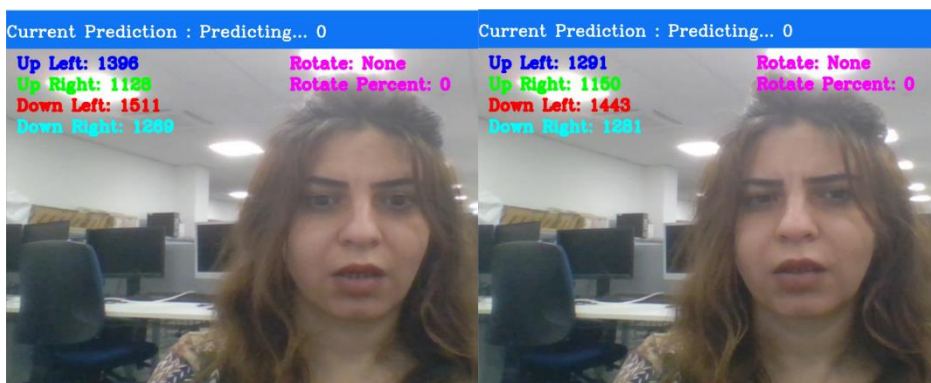


Figure 5.11 Other FEs examined by the model and the model shows (0) value for predicting which means the stability medical condition of patient.

5.4 Support Vector Machine (SVM)

Figure 5.12 presents the confusion matrix that summarises and visualises the performance of the proposed model. Each row of the matrix presents FEs in the actual class, while each column represents FEs in the predicted class.

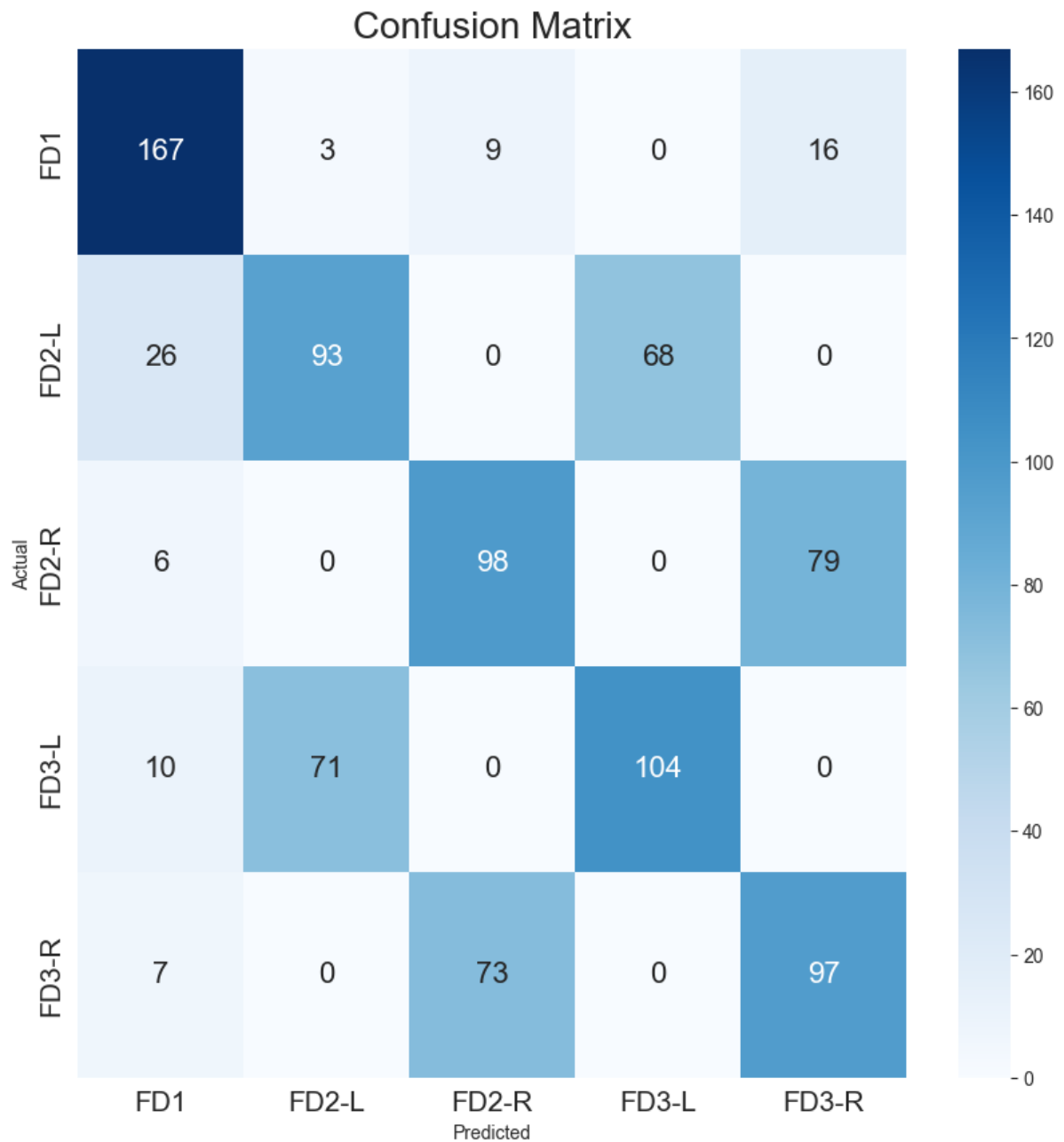


Figure 5.12 The Confusion Matrix of SVM Model.

Figure 5.13 & Figure 5.14 illustrate all evaluated measurements of the proposed model.

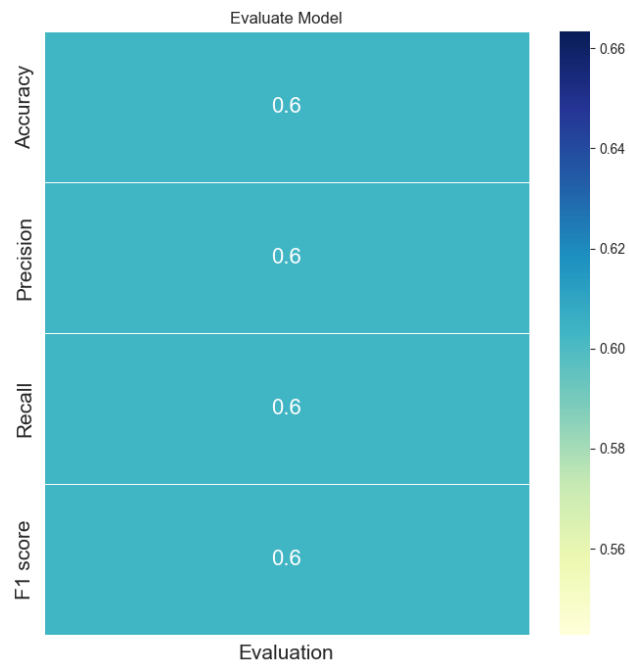


Figure 5.13 The Evaluation Metrics of SVM Model.

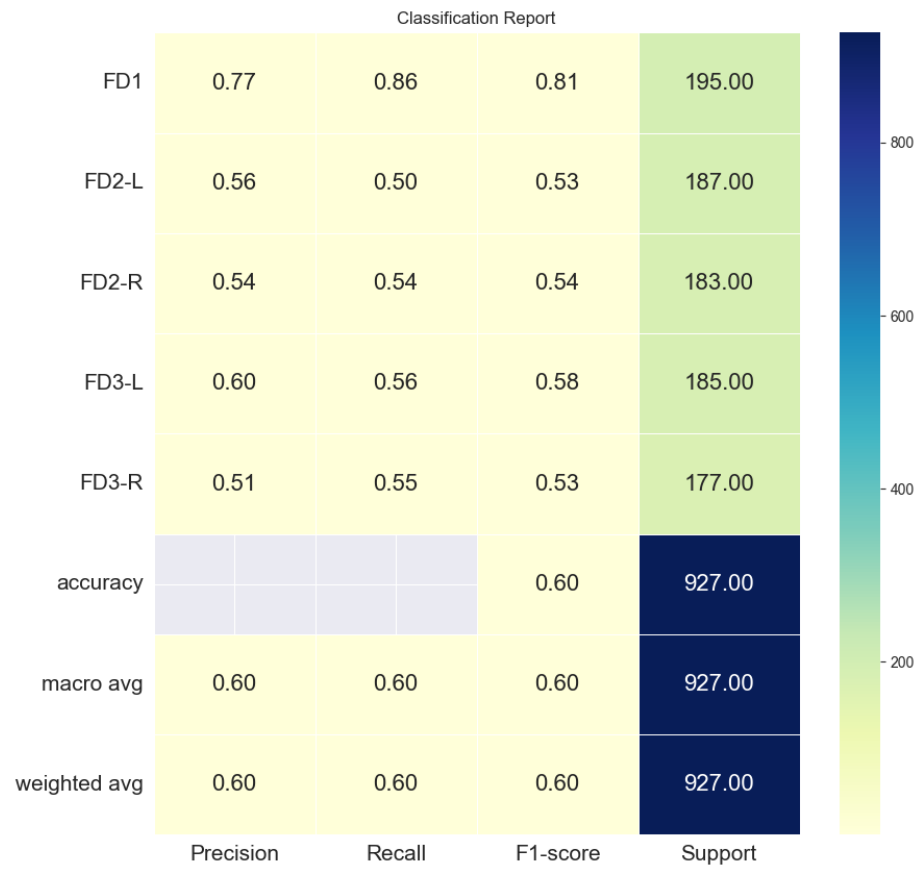


Figure 5.14 The Classification Report of SVM Model.

The five evaluation measurements for five classes are illustrated in the following Table 5.2:

Table 5.2 Evaluation metrics for each Class & The Total Mean Performance of SVM Model.

Class Name	Facial Expression	Precision	Recall	F1 Score
FD1	AU (15+43+25)	79%	84%	82%
FD2-R	AU (15+43+55)	52%	57%	55%
FD2-L	AU (15+43+56)	56%	47%	51%
FD3-R	AU (15+25+43+55)	51%	49%	50%
FD3-L	AU (15+25+43+56)	58%	61%	59%

The above measurements provide insights into various aspects of model performance, including the precision, recall, and F1 score. The model accuracy is 0.597 (or 59.8%) which is calculated by the formula provided in Chapter 3.

5.5 Conclusion

In this chapter, we explored the classification of FEs using RF and SVM algorithms to identify facial expressions of health deterioration, categorized into five distinct classes: FD1, FD2-R, FD2-L, FD3-R, and FD3-L.

Our proposed method involved several stages, starting with data generation through 3D facial models and deepfake techniques to create realistic training samples as highlighted in chapter 4. We utilized various pre-processing techniques, including resizing images, converting colour spaces, and extracting 3D facial landmarks using MediaPipe's Face Mesh algorithm. The key to our feature extraction method was leveraging the geometric features derived from the facial landmarks, which provided a detailed representation of facial movements and expressions.

For the classification task, we employed the Random Forest algorithm, which demonstrated exceptional performance, achieving a state-of-the-art accuracy of 100%. This was validated through comprehensive evaluation metrics such as precision, recall, and F1 score, all of which

also recorded 100%. These results underscore the robustness of our feature extraction and classification approach.

The model is tested in real-time scenarios using the author's facial expressions and showed significant accuracy, it faced challenges with transient expressions and environmental variables, highlighting areas for potential improvement in real-time applications.

We also compared the performance of the SVM classifier, which achieved an accuracy of 59.8%. Although this is lower than the RF model, it provides valuable insights into the different capabilities and limitations of various machine learning techniques in facial expression recognition.

Overall, the results from both classifiers indicate that the integration of MediaPipe for feature extraction and RF for classification provides a powerful solution for real-time facial expression recognition. Future work can focus on improving real-time performance to further refine the model's applicability in medical condition monitoring.

Chapter 6

Facial Expression Classification Using Deep Learning

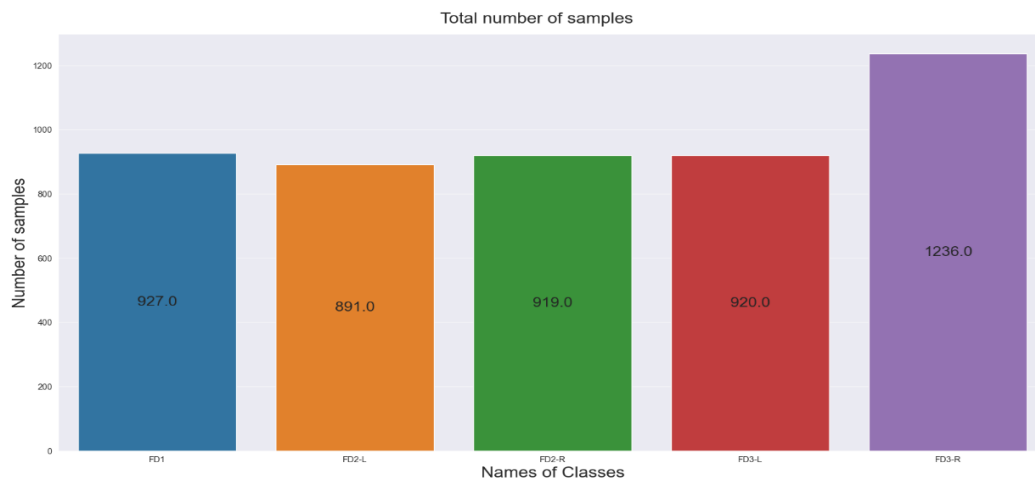
6.1 The Proposed Method using Deep learning Models

Deep Learning (DL) models do not need any feature extraction due to their ability to read and analyse macro and micro expressions and learn the hierarchical representations of features directly from raw data. The aim of proposing various AI algorithms and different preprocessing techniques is to identify optimal method combinations that can predict whether the video frame contains a specific combination of AUs in the patient's face.

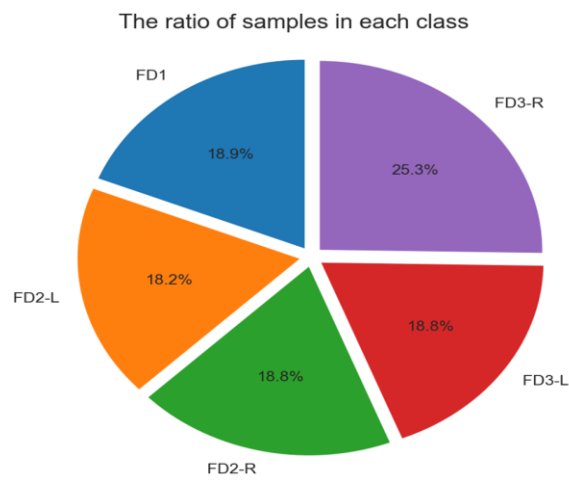
The first proposed model is based on a CNN algorithm to identify FE patterns, and a second model is comprised of a hybrid of convolutional and recurrent memory neural networks. The two models are implemented using Keras and TensorFlow which are open-source software libraries for artificial neural networks. Keras is an API designed to be used to act as an interface, while TensorFlow is an end-to-end framework suited for dataflow programming. Finally, a third model is based on Transformer networks. Training the DL models using generated data optimises the models to develop a system that can achieve high standardisation and precise outcomes.

Before feeding the raw data to DNN models, detecting and identifying faces is considered a crucial step in the AFER pipeline, providing relative data by focusing on the face region of interest to capture detailed information on FEs and to classify their types. Faces provide a wealth of information through FEs, including the position, intensity, and appearance of AUs for specific facial muscle movements. Therefore, providing a model with the region of interest by localising and aligning the face area aids in isolating the face from the background and identifying the relevant features of interest. This reduces the impact of irrelevant data like background and noise that may affect the accuracy of model prediction. Furthermore, unnecessary computations that may reduce model performance are avoided.

The videos generated were divided into five classes according to their face display (FD). The total number of samples of each class are illustrated in Figure 6.1, revealing a slightly imbalanced dataset.



(a)



(b)

Figure 6.1 Number and ratio of Samples in each class for the whole dataset.

(a) The total number of Samples is represented by Column Chart.

(b) The ratio of Samples in Each Class

The whole dataset was then split into 85% training and 15% test datasets. Test data is essential to evaluate model performance on unseen data. It is worth noting that any model performs more precisely when it is fed with a rich, sufficient, and diverse dataset. Figure 6.2 provides the number of samples for both training and testing datasets.



(a)



(b)

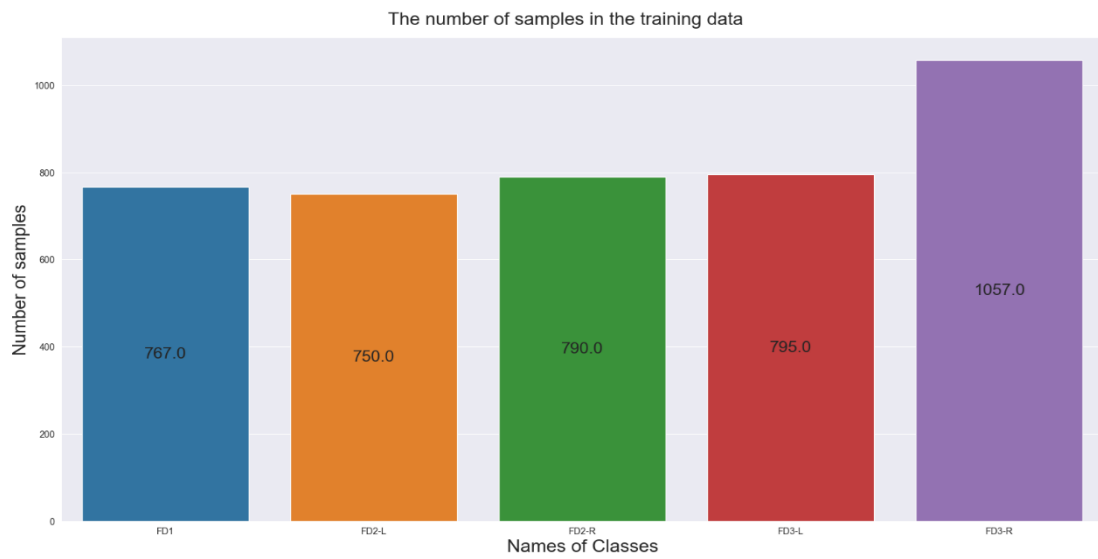
Figure 6.2 Number of Samples in training and test dataset.

(a) Number of samples in training dataset.

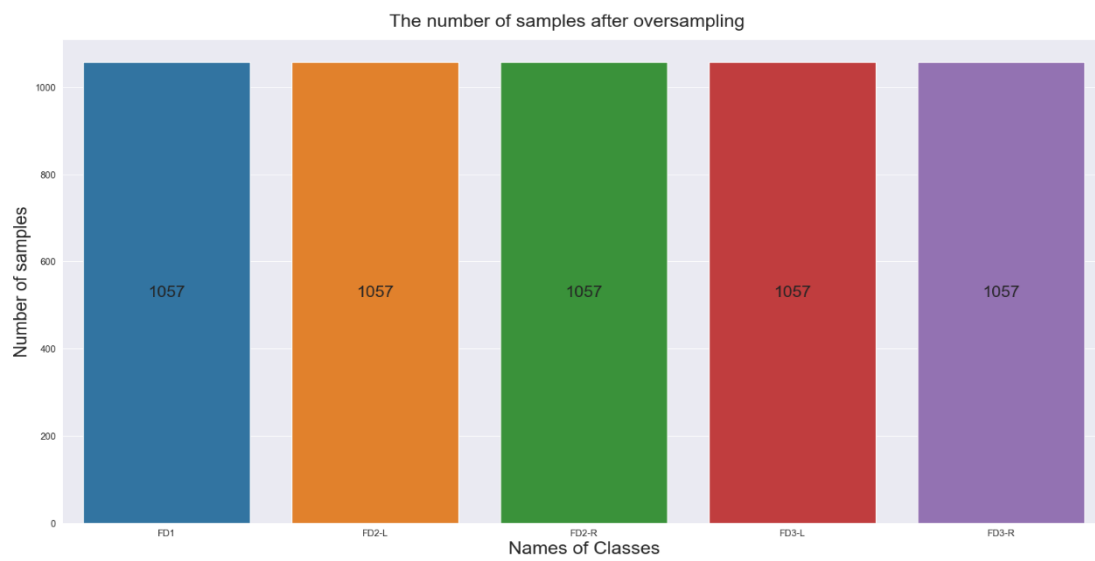
(b) Number of samples in test dataset.

The final step in preprocessing is oversampling, which is effective in ANNs for the handling of imbalanced classes and to improve model performance by providing it a balanced training set. The oversampling method was only applied to training datasets to avoid data leakage. Evaluating a model's performance on an imbalanced test dataset is crucial to assess its ability to perform real world generalisation.

Figure 6.3. depicts the training dataset before and after oversampling.



(a)



(b)

Figure 6.3 Number of Training Samples Before and After Oversampling Method.

(a) Number of Samples of Training Dataset Before Oversampling.

(b) Number of Training Dataset After Oversampling.

Figure 6.4. depicts the method of automatic facial expression recognition based on Deep Learning Neural Networks.

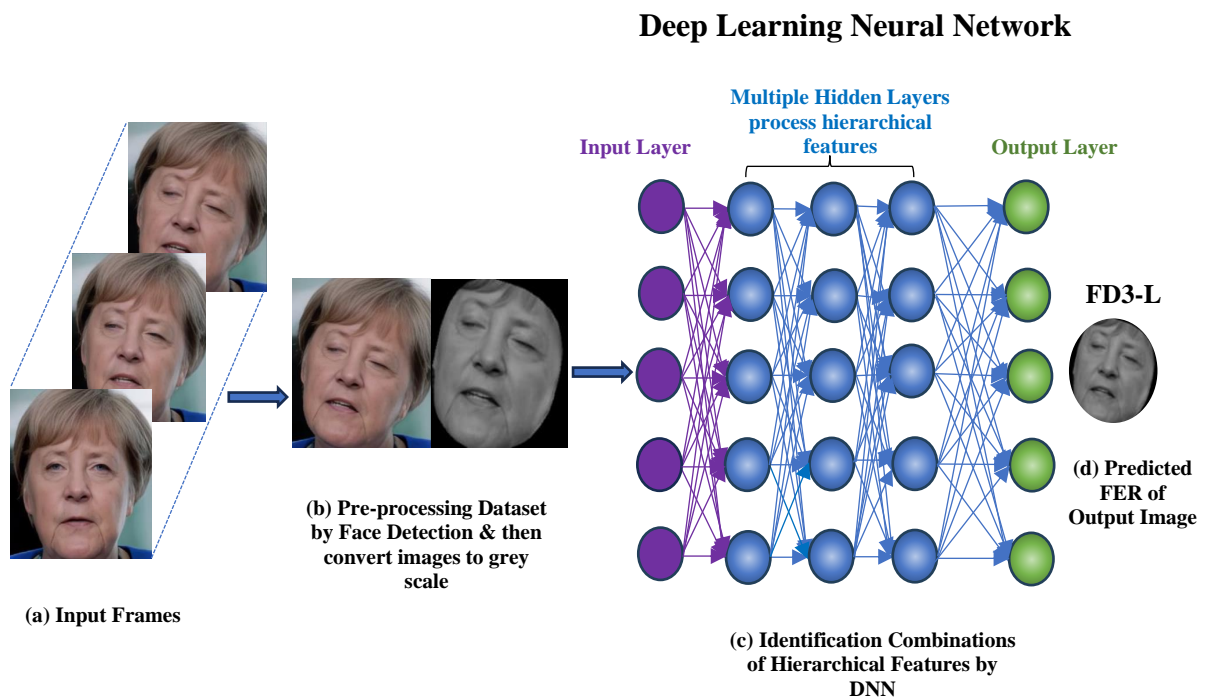


Figure 6.4 Phases of AFER based on DNNs. (a) Input Frames. (b)Pre-processing frames using Face Detection and convert them to grey scale. (c) Identifying hierarchical spatial and temporal FEs features. (d) a single face expression is recognized and categorized. (face images are taken from generated dataset called PRD-FE).

6.2 A 1D-CNN Model to Predict Patients at Risk of Deterioration

CNN is a deep learning method widely adopted in image analysis and classification problems due to its remarkable performance in detecting and recognising patterns in images. CNNs can be used to solve various problems in applications such as self-driving cars, object recognition, movement detection, and FER. Recent advances in computer vision indicate that the most desirable approach for obtaining image characteristics is a CNN architecture because of its ability to produce high accuracies in image recognition.

In a CNN, a combination of features can be automatically extracted to analyse, identify, and classify each image. CNN obtains image characteristics by using convolution layers. These layers apply a transformation to the image called image convolution. Each convolution layer extracts feature vectors and is different from the next layer. Each single image contains various

patterns such as textures, edges, objects, and shapes, and different convolution filters can detect these patterns selectively. The deeper layers use more sophisticated filters that can detect more specific features. The matrix of each filter is initialised with random numbers and the filter will slide over each set of pixels from the input image implementing a convolution process to detect patterns.

After preprocessing, the next step is designing and evaluating the model. This phase is composed of three stages: the first is building the deep learning model, the second is training the dataset, and the third plots the performance for evaluation.

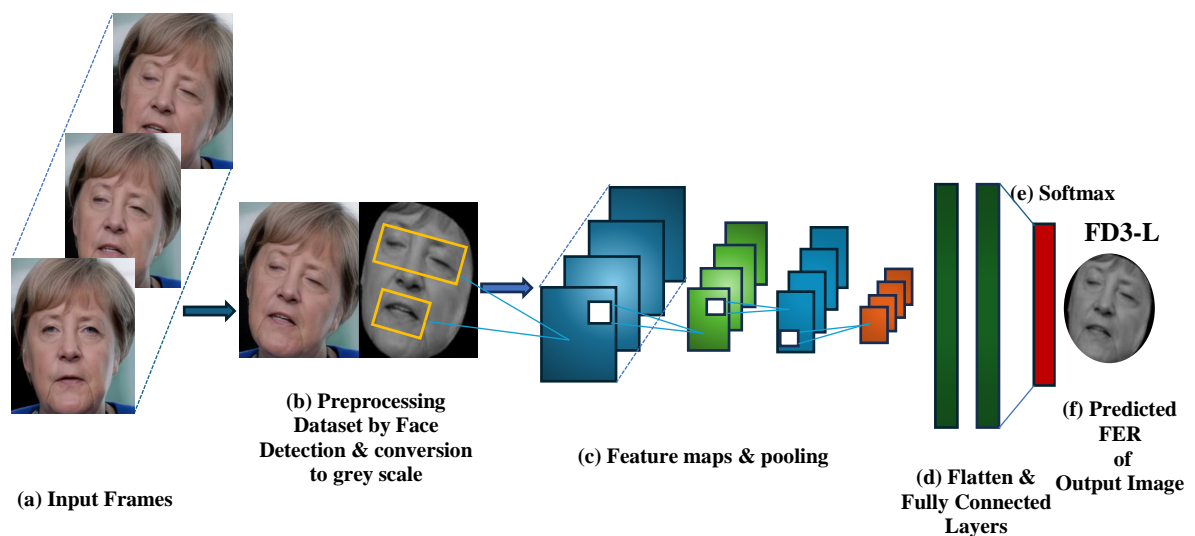


Figure 6.5 CNN-based FER method. (a) Input Frames. (b) Pre-processing data by Face Detection and conversion to grey scale. (c) From the convolution results, feature maps are constructed, and max-pooling (subsampling) layers lower the spatial resolution of the given feature maps. (d) Flatten & Fully Connected Layers, and (e) a single face expression is recognized based on the output of softmax (face images are taken from generated dataset PRD-FE).

The proposed model in this thesis combines face detection techniques with 1D-CNN, which extracts feature patterns from the generated dataset representing AUs in the upper and lower parts of the face. The combined model can adapt to differences in face location according to the camera and environment, and differences in facial landmarks, age, skin tone and ethnicity. The model was tested on unseen data representing these variables and evaluated using k-fold cross-validation for each participant individually. One round for each participant was held for testing and the rest were used for training. The model accuracy was the average accuracy of all rounds.

The FE prediction model consists of two main parts, namely the face detection and the DL algorithm. To validate the 1D-CNN, two other DL models were also used (ConvLSTM and Transformers) to evaluate their performance. The best performance was observed for the 1-CNN due to its ability to capture spatial features from raw input data, which enable prediction of a patient's deterioration with a high degree of accuracy. The method is illustrated in Figure 6.5.

6.2.1 Data Pre-processing and Design of a 1D-CNN Model

Data Pre-processing

The preprocessing methods significantly improve the performance of the learning process and model generalisation by enhancing the quality of the dataset, minimising noise, introducing variability, and providing standardised input data for ML models. The appropriate method is selected based on the nature and characteristics of the data, the requirements of the DL model, and the target task. It is crucial to achieve a balance between increasing the variability and preserving the essential features of the dataset. This section presents the preprocessing methods employed for the proposed system based on CNN.

The first step after reading the video streams is labelling all videos according to their classes. At the early stage of preprocessing, the read frames will be converted from BGR to RGB images, then a face detection model is applied for each frame, deleting the background to avoid unnecessary computation, noise and unwanted data. For real-time processing, the frames are captured from the camera and submitted to the same preprocessing method.

The advantage of the face detection model provided by MediaPipe is its high accuracy due to the DL algorithm trained on a large dataset and fine-tuned for optimal performance. The model can accurately detect faces and FEs through images and a continuous stream of images under various conditions of lighting and orientations. This method is also convenient in real-time systems due to high speed, achieved by leveraging efficient algorithms and hardware acceleration such as GPU. Moreover, this model is designed to be robust to various factors that can affect face detection accuracy, such as occlusions (e.g., partial face coverage by objects or hands), varying distance from the camera, and different camera angles. This robustness ensures reliable performance across a wide range of scenarios and internal and external factors.

To achieve consistency with MediaPipe across various image datasets, it is crucial to resize images to a desired size. In this thesis, the input images were normalised to a fixed width and

height of 224*224 pixels for improved performance. For image process tasks and computer vision, the normalisation method is commonly combined with other preprocessing methods such as cropping, and data augmentation. Hence, after resizing and flattening the images or video frames, the normalization approach is the next step to improve performance of the learning process. Therefore, scale and standardization are applied to the features by dividing them by 255 and storing them in a standard NumPy array. Due to the scale sensitivity of ML algorithms, this method normalises feature magnitudes, avoiding bias towards those of higher magnitude that will dominate the learning process with a negative impact on the performance and convergence of DL algorithms. In addition, the convergence of algorithms (based on gradient descent) may struggle for un-normalised scales, requiring longer processing times or failing to achieve the optimal outcome.



Figure 6.6 Preprocessing Dataset by Face Detection & then convert images to grey scale.

Converting colour images to grayscale can also reduce computational complexity while still allowing the recognition of relevant facial patterns. Furthermore, involving face detection techniques can help DL algorithms by avoiding irrelevant data like the background or the human body. Figure 6.6 shows the converted grey scale image before and after face detection and face oval techniques.

Design of a 1-D CNN

Input images are convolved into a feature or activation map and passed to the next convolution layer. Every single neuron will connect to another single neuron to produce a fully connected layer. Figure 6.7 presents the proposed CNN architecture using the generated dataset PRD-FE. The proposed 1D-CNN architecture processes sequential data like time-series data, audio signals, or any data that can be represented in a one-dimensional format.

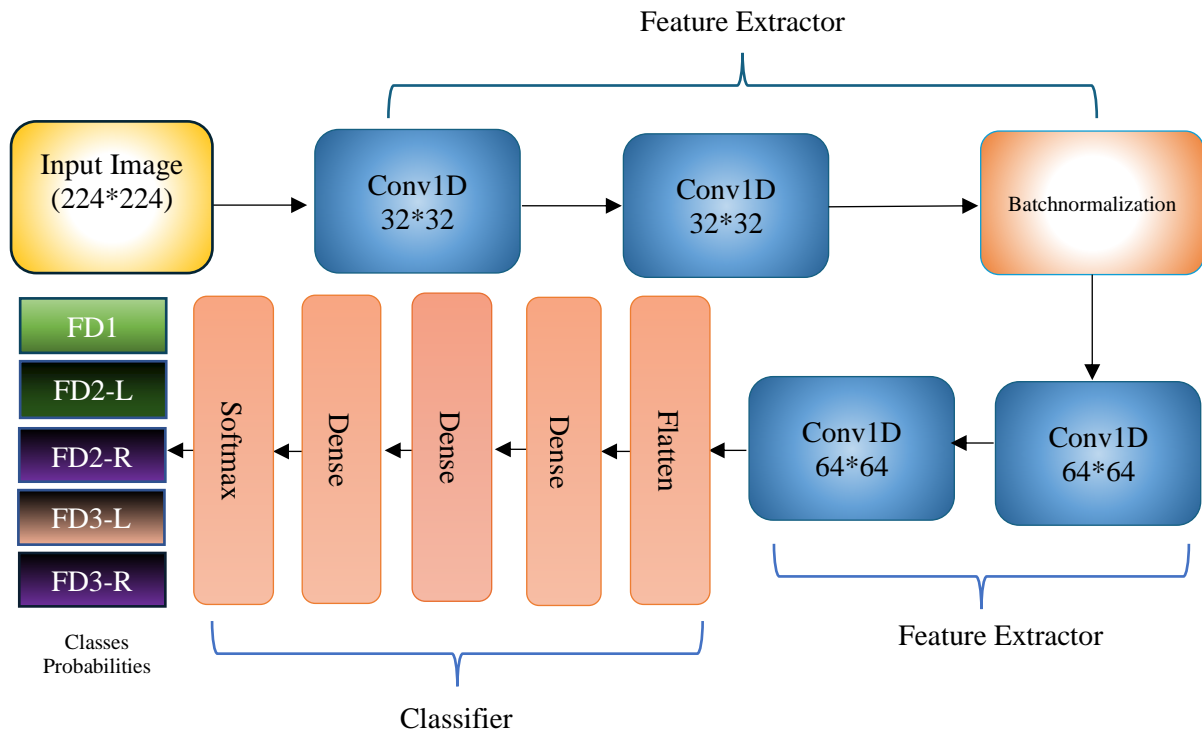


Figure 6.7 The design of the 1-D CNN model.

A breakdown of each component and its function within the proposed model is described as follows.

- **Input layer.** Accepts the input with the shape (None, 5, 224, 224). This indicates a batch of data with 5 time steps, along with the image of size 224*224 pixels.
- **1D convolution layers (Conv1D with ReLU activation).** These layers apply convolution operations along the time dimension, reducing feature space dimensions (from 224 to 218) while transforming the data. The number of filters is increased from 32 to 64, enabling the network to capture more complex features in the data.
- **Batch normalisation.** Normalises the output of the convolution layers, stabilising learning by reducing internal covariate shift.
- **Flatten layer.** Converts the multidimensional output of the convolution layers into a 1D array, making it suitable for input into the fully connected layers.
- **Dense layers.** These are fully connected layers that further process features extracted by the convolutions. The network reduces the dimensionality progressively through these layers (from 128 to 32), focusing on the most relevant features.

- **Final dense layer (dense with Softmax activation).** This layer has 5 units corresponding to the output classes, with Softmax activation to generate a probability distribution over these classes, indicating the likelihood of each class being the correct classification for the input.

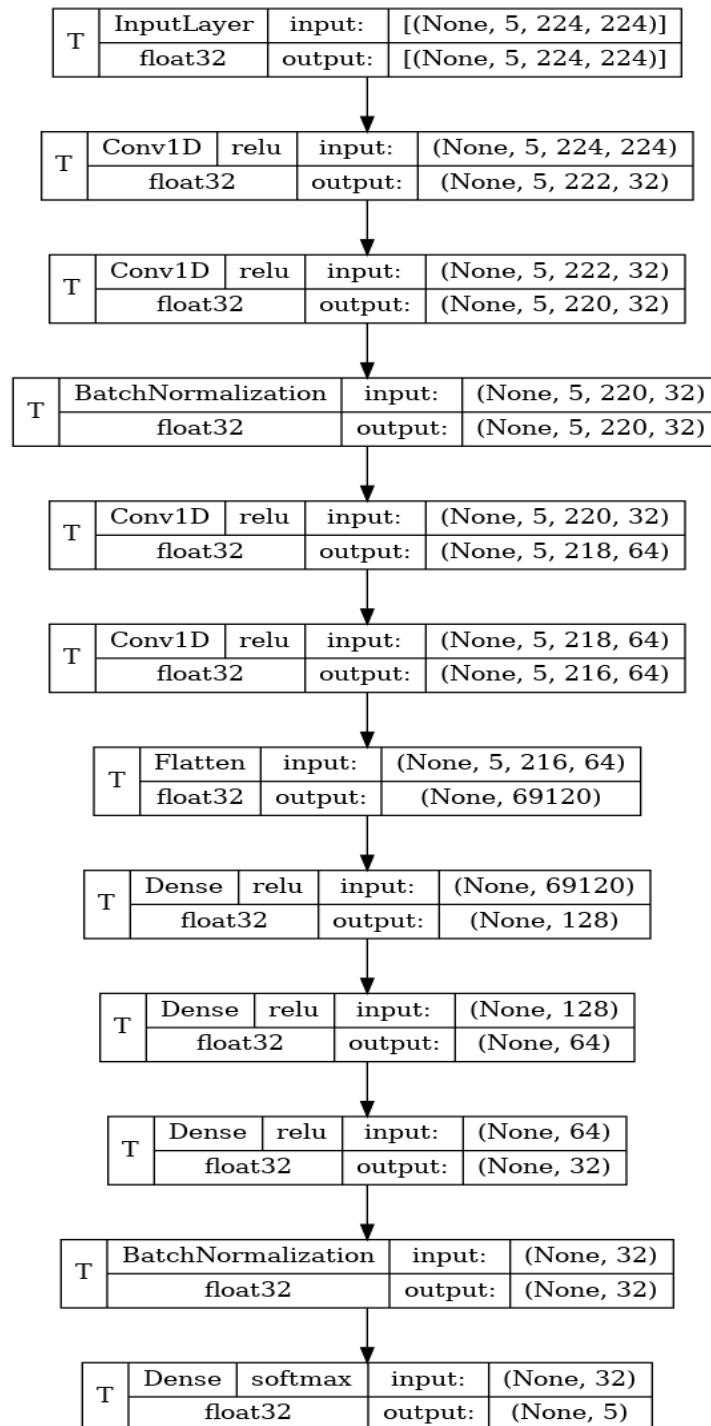


Figure 6.8 Layers of proposed 1D-CNN model.

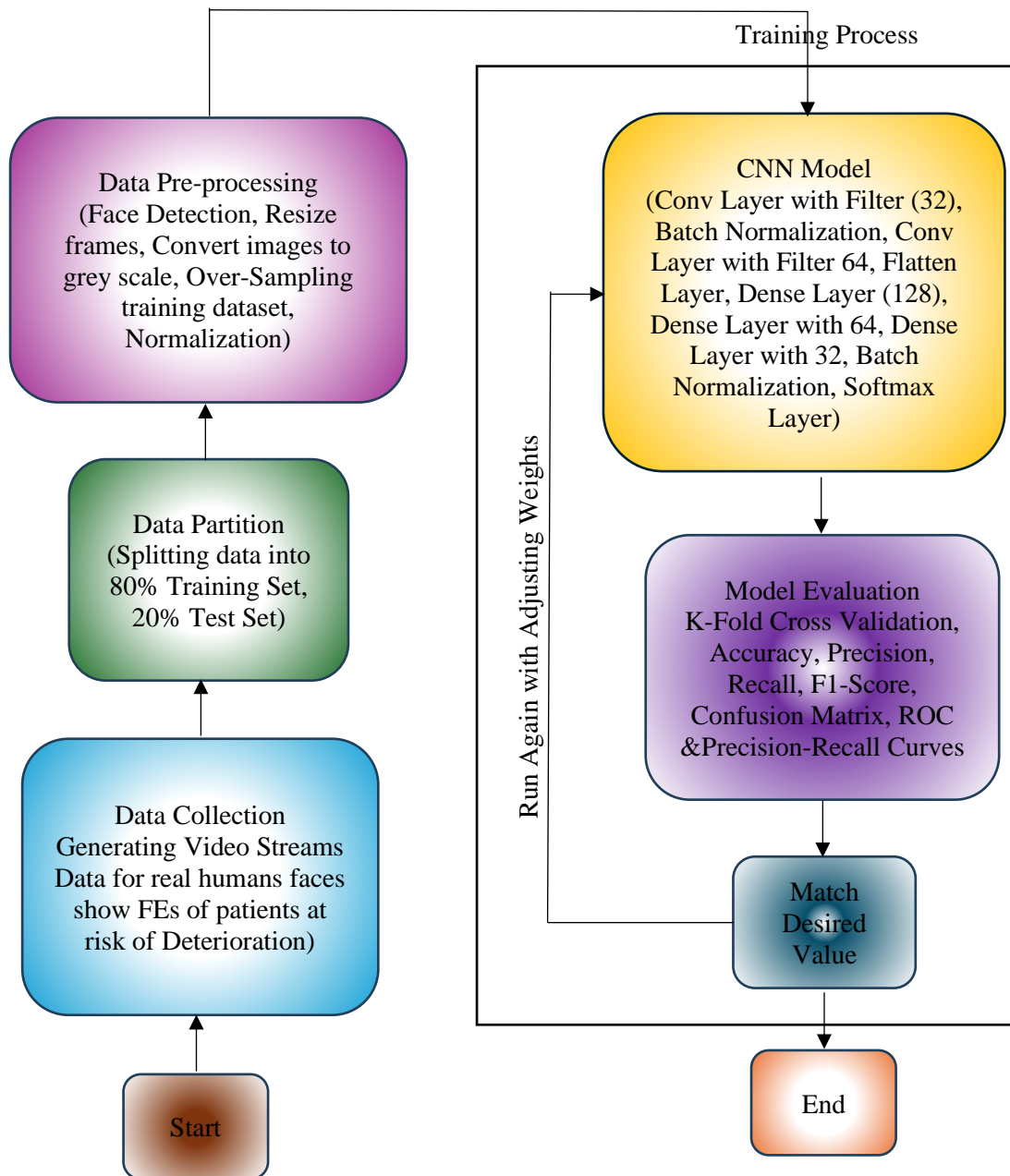


Figure 6.9 The stages of 1D-CNN based FER.

This architecture is structured to gradually refine and abstract the input data features, culminating in a classification output. Each layer's specific purpose is to transform the data progressively into a format optimal for the classification task at hand. Figure 6.8 depicts details of the proposed model layers and Figure 6.9 illustrates the stages of the FER-based 1D-CNN model.

6.2.2 Results and Evaluation of the 1D-CNN Model

The proposed model was applied to the generated dataset PRD-FE. The dataset includes video frames for the 5 classes of FEs with various facial landmarks, skin tones, and ethnicities. Optimal results depend on many factors such as the quality, quantity, and diversity of the dataset, the effectiveness of feature extraction methods, the model structure, experimentations, and fine-tuning.

The model was evaluated by calculating the precision, recall, and F1-score of the training and testing dataset along with the confusion matrix. Precision is used to measure the number of correctly predicted samples of positive class. It is calculated as the ratio of correct sample predictions to the overall number of samples identified as positive class. Figures 6.10 and 6.11 show the accuracy and loss of the proposed model for the testing dataset.

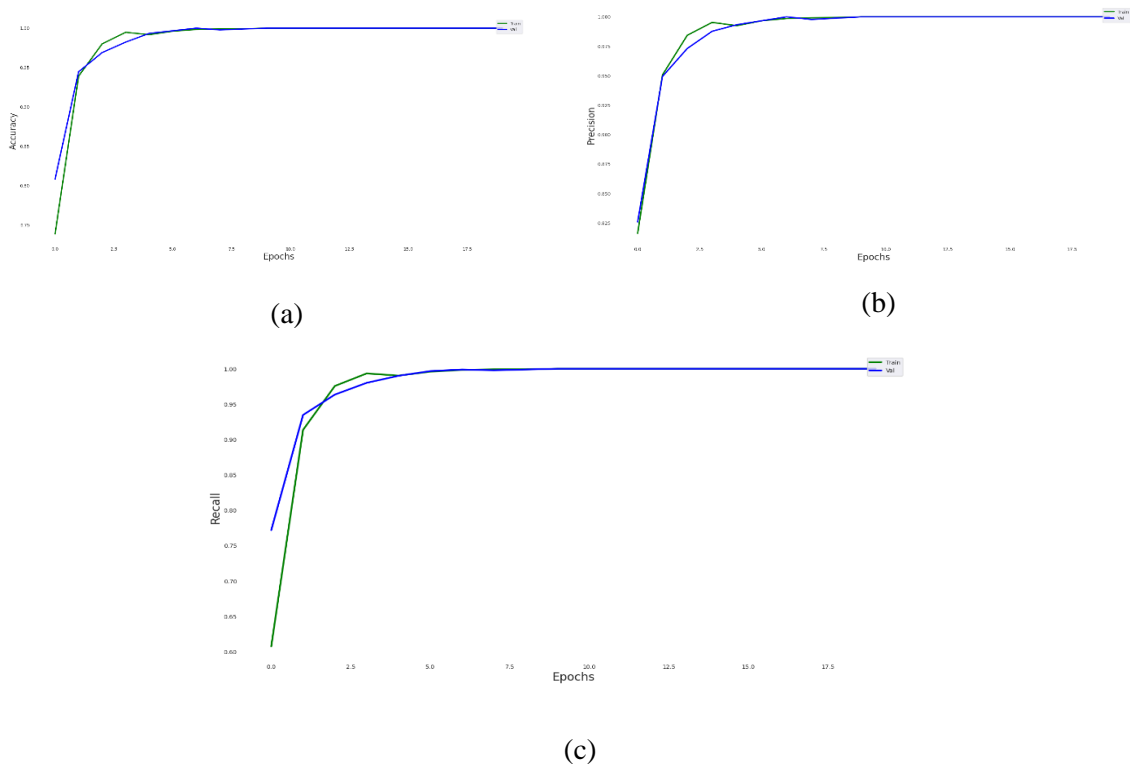


Figure 6.10 Evaluation Metrics of 1D-CNN Model Performance. (a) Accuracy of Proposed Model. (b) Precision of Proposed Model. (c) Recall of Proposed Model.

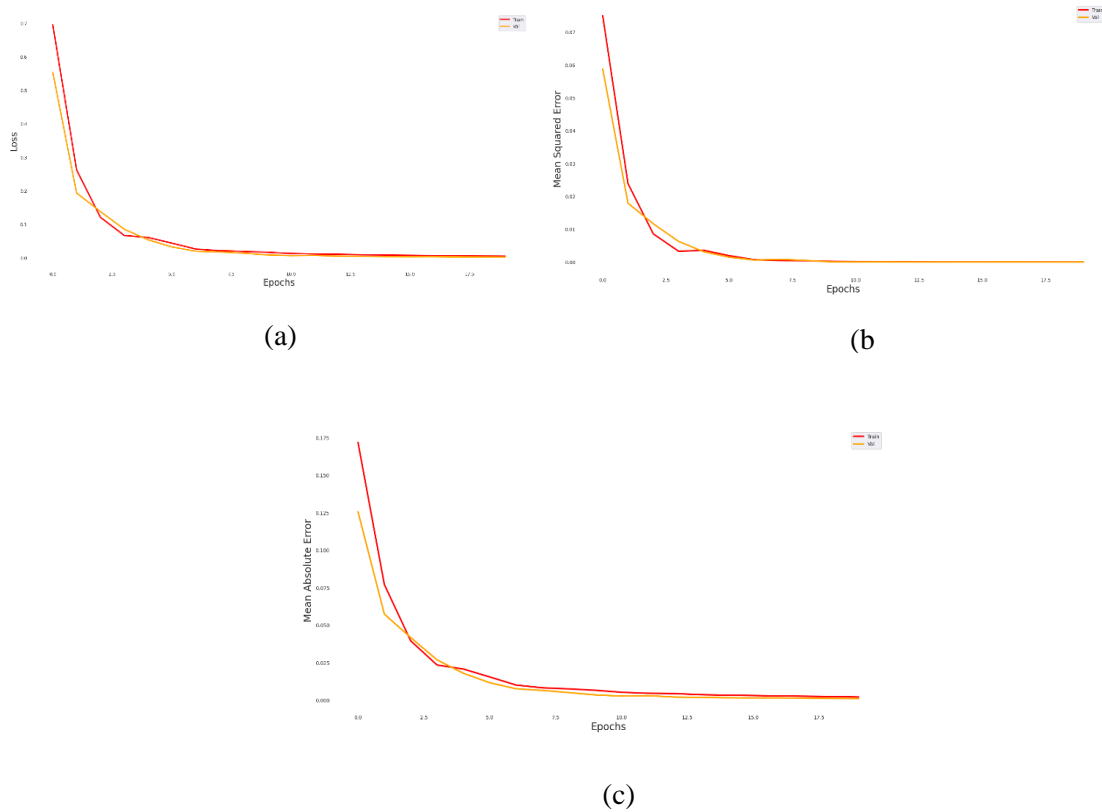


Figure 6.11 Loss, Mean Square Error, and Mean Absolute Error of 1D-CNN model. (a) Loss of Predicted Model. (b) Mean Square Error of Predicted Model. Mean Absolute Error.

Figure 6.12 presents the confusion matrix that summarises and visualises the performance of the proposed model. The structure of the confusion matrix includes rows that represent FEs in the true classes (FD1, FD2-L, FD2-R, FD3-L, and FD3-R), while columns represent FEs in the predicted classes (FD1, FD2-L, FD2-R, FD3-R, and FD3-L). The analysis of each face display (FD) in the confusion matrix for the CNN model is as follows (TP = true positive, FP = false positive).

- For FD1 (row 5), 166 TP samples accurately identified as FD1, and 0 FP samples misclassified into other categories.
- For FD2-L (row 2), 167 TP samples accurately identified as FD2-L, and 0 FP samples misclassified into other categories.
- For FD2-R (row 4), 142 TP and 0 FP.
- For FD3-L (row 1), 169 TP but 2 FP samples misclassified as FD2-L.
- For FD3-R (row 3), 151 TP and 0 FP.

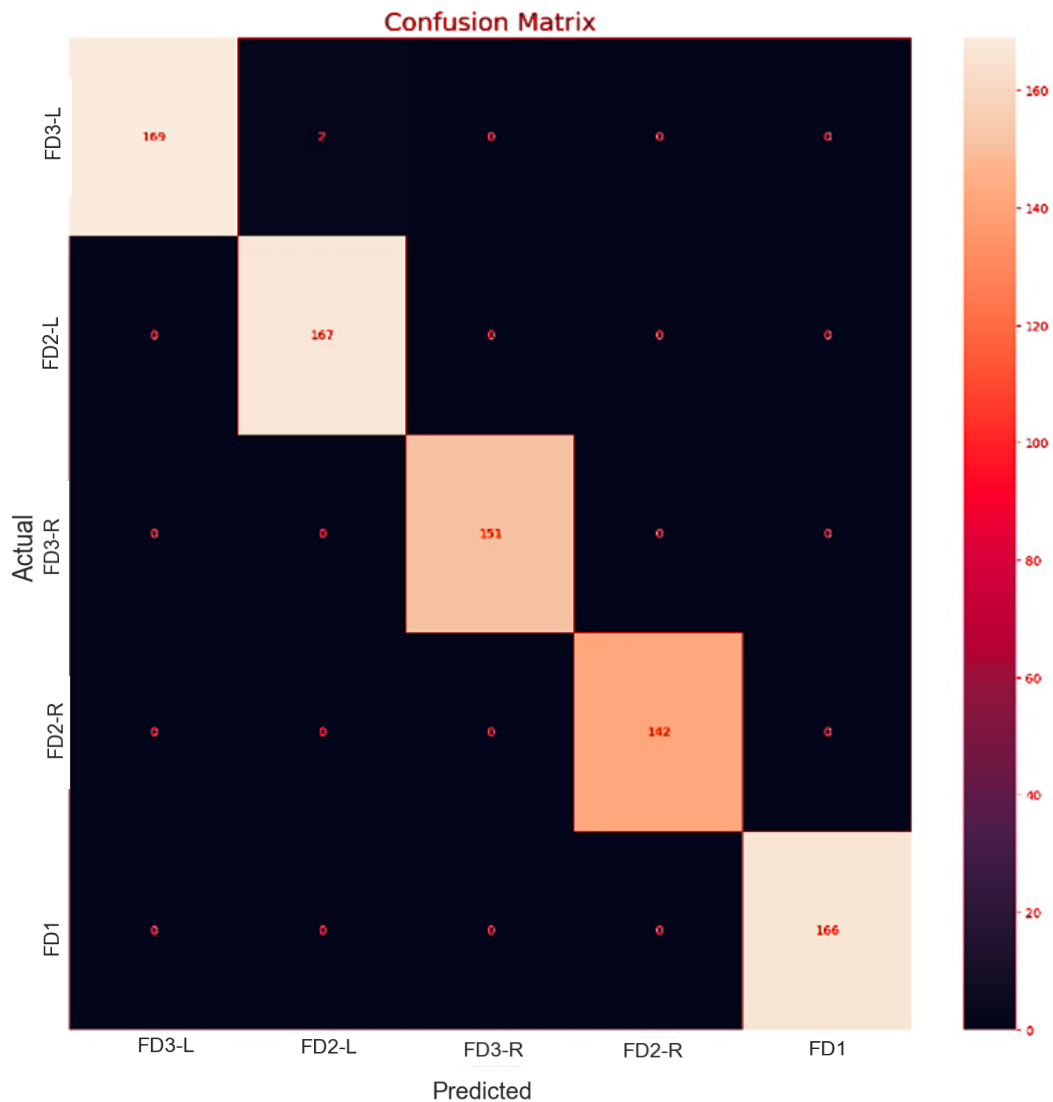


Figure 6.12 The Confusion Matrix of 1D-CNN Model.

Hence, the model has a high degree of accuracy in classifying all classes, with most instances in each category correctly classified. There is a very low misclassification rate, with only 2 samples of FD3-L misclassified as FD2-L.

In summary, the model shows balanced performance across the 5 FD classes, with no single class showing a significantly higher error rate than another. The model has effectively learned to differentiate between the features of different classes. Moreover, the model's ability to generalise across multiple classes without significant bias toward any class is evidence of proper training and diversity in the training data.

Although the misclassification rate is low, it is useful to examine the reason for this confusion to further refine the model's accuracy. Further validation with unseen datasets would ensure robustness and the ability to generalise beyond the training data.

Figures 6.13 and 6.14 depicts all evaluated performance of the proposed model.

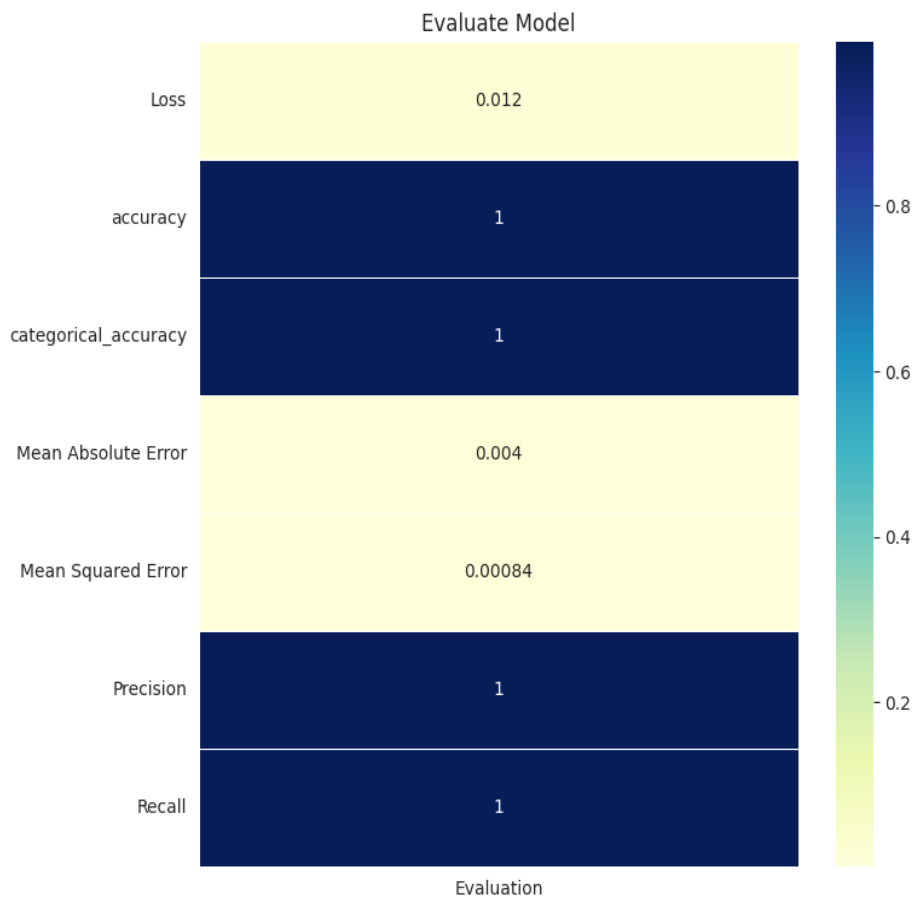


Figure 6.13 The Evaluation Metrics of 1D-CNN Model.

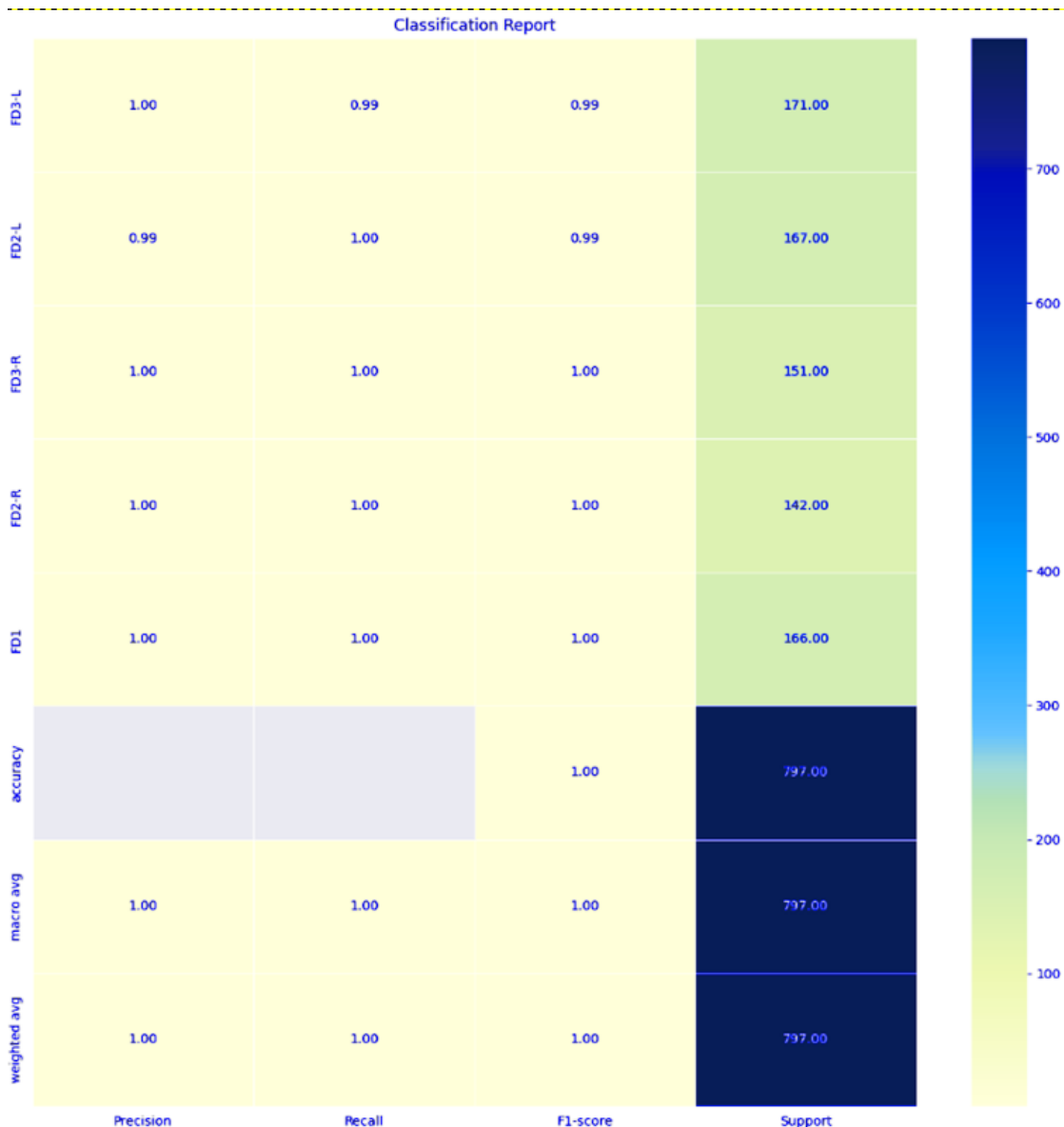


Figure 6.14 The Classification Report of 1D-CNN model.

According to Figure 6.14, all classes FD1, FD2-L, FD2-R, FD3-L, and FD3-R show an area under the curve (AUC) of 1.00. This indicates that the model shows perfect classification capability between the positive and negative classes for each category without error. Each curve reaches the top-left corner of the plot (TPR = 1, FPR = 0), which is the ideal point for a classifier, indicating 100% sensitivity (no false negatives) and 100% specificity (no false positives) at a certain threshold.

As the model achieved an AUC of 1.00, which is rare in real-world scenarios on unseen data, two possible conclusions can be drawn. As the first possibility, the dataset is controlled and highly specific, allowing the model to perfectly learn the distinctive features and categorise samples with 100% sensitivity and specificity at a certain threshold. In practical terms, this

means that there are no false positives or false negatives at this optimal threshold. As the second possibility, the model is overfitting the training dataset, meaning that it may not perform as well on unseen dataset. This is not the case for the proposed model as it performs well on an unseen dataset.

Such perfect results might indicate that the model is overfitting the data from which it was developed, unless this is a controlled, or this is a very specific dataset where perfect classification is achievable. Let us discuss the first reason which is the possibility of generated dataset with highly controlled conditions such as minimizing the variation of images lighting, face alignment, and images backgrounds, which is not the case for our generated database as it created with unlimited conditions.

For the generated dataset, the images produced with various lightings degrees, face alignment (scale, rotation), different backgrounds as highlighted in chapter 4. For face alignment various techniques have been experimented as one of the most important requirements for producing predicted model with high precision and performance is eliminating noisy and irrelevant data from the raw dataset. Therefore, the MediaPipe face detection and face oval have been utilized as preprocessing techniques that contribute to eliminate irrelevant data, such as image background. This is necessary as it may result in the DNN model to learn these irrelevant data and consider them as discriminative features and associate those background features with specific outcomes and that may lead to model failure with real world unseen data. Overall, as the generated dataset has been generated with various conditions and the model achieved significant classification precision for unseen data, it means the ideal AUC is not due to overfitting problem but due to the model that has been trained properly with optimal parameters and hyperparameters (learning rate, batch size, etc.), and architecture (number of layers, activation function). If these results are reproducible in real-world conditions outside of the test dataset, the classifier would be exceptionally effective. However, it is important to validate these findings with external datasets to ensure that the model generalizes well and is not simply tuned to the data used for this analysis.

The last probability is the dataset has not been used with real-world data and this is true as it is simulated data; however, it is generated with various conditions (lighting, face scale, face rotation, images' backgrounds) and mimic as close as possible the real human FEs using various intensity of AUs for different face shapes, facial landmarks, skin tone, age, and ethnicity. In future work, it is crucial to collect real samples data in a clinical scenario and

carried out further tests to ensure robustness of the proposed model before real world practical deployment.

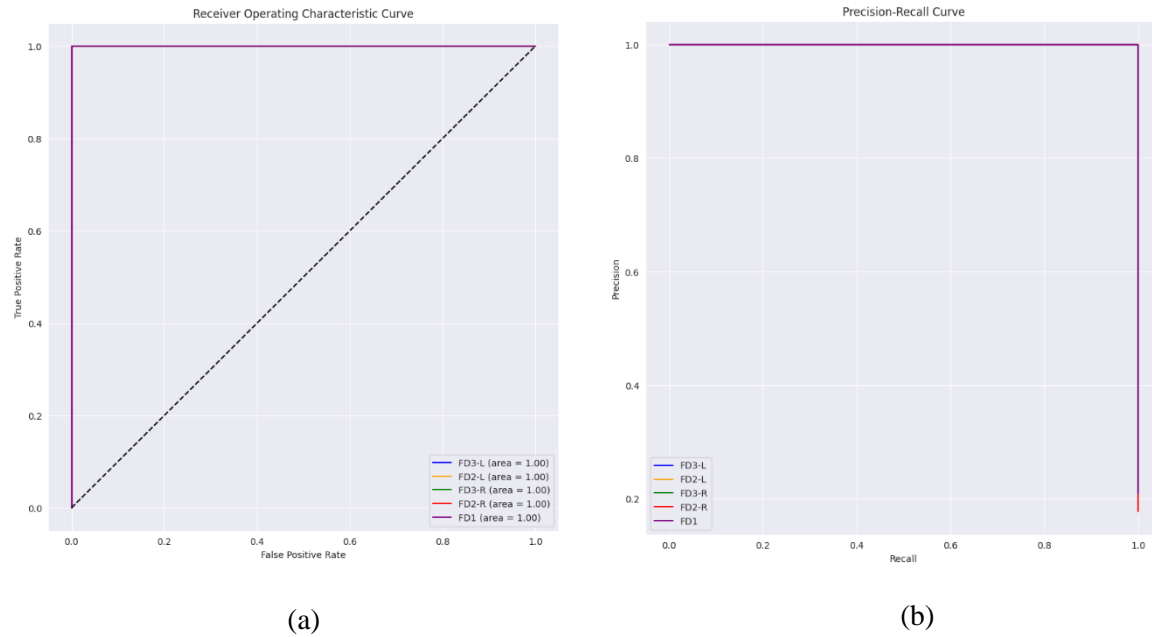


Figure 6.15 Evaluating 1D-CNN Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve. (a) ROC. (b) Precision-Recall Curve.

Figure 6.15a shows the ROC curve and Figure 6.15b the precision-recall (PR) curve, which are used to evaluate the performance of a classifier, especially when dealing with imbalanced datasets. Each line represents the trade-off between precision (y-axis) and recall (x-axis) for a different class (FD1, FD2-L, FD2-R, FD3-L, and FD3-R).

Precision measures the accuracy of the positive predictions and is calculated as the ratio of true positives to the sum of true and false positives, as described in Chapter 3. Recall (sensitivity) measures the ability of the model to find all the true positives and is calculated as the ratio of true positives to the sum of true positives and false negatives, also described in Chapter 3.

The curves for all classes are located at the top-right corner, indicating both high precision and high recall with values close to 1.0 (the model correctly classifies almost all positive examples without incorrectly labelling negative examples as positives). High precision and high recall indicate that the model is both accurate and comprehensive in its positive classifications. Such performance may be due to the controlled data generated under ideal conditions, or very distinct classes that are easily distinguishable by the model. As the PRD-FE database is generated with various conditions of lighting, face scale, face rotation, and background, the data and model must be exceptionally well-suited to the task.

In terms of validation prior to real world deployment, ideal scores may indicate overfitting, meaning that the model is too closely fit to the training data and may not generalise well to unseen data. Alternatively, the training conditions or data might be ideal (e.g., very clear distinctions between classes, high-quality annotations, minimal noise). Validating the model with independent test data not seen during the model training will confirm the model’s ability to generalise with high accuracy predictions, meaning that the model is well optimised and trained for diverse datasets. Therefore, the model must be tested with real human faces varying in age, gender, skin tone, ethnicity, and image background. Since it was not possible to conduct a full clinical trial during this PhD research, we emulate this process using real-world images of public figures like Tom Hanks, Johnny Depp, and Angela Merkel.

Early models designed and tested during this research project were too complex and suffered from overfitting, despite the approximately 96% accuracy for unseen data. As the model failed to generalise, it was simplified using techniques such as regularisation. Extensive fine-tuning of model parameters gave the ideal performance shown by the PR curve. The five evaluation measures used in this thesis are summarised in Table 6.1 for the 5 classes of face displays under investigation.

Table 6.1 Evaluation Metrics for each Class % The Total Mean Performance 1D-CNN Model.

Class Name	Facial Expression	Precision	Recall	F1 Score	Accuracy
FD1	AU (15+43+25)	100%	100%	100%	100%
FD2-R	AU (15+43+55)	100%	100%	100%	100%
FD2-L	AU (15+43+56)	99%	100%	99%	100%
FD3-R	AU (15+25+43+55)	100%	100%	100%	100%
FD3-L	AU (15+25+43+56)	100%	99%	99%	99%
	Mean	99.8%	99.8%	99.6%	99.8%

The above measurements provide insights into various aspects of model performance which recorded 99.8% precision, 99.8% recall, 99.6% F1 score, and 99.8% accuracy. The model was also evaluated on unseen data, showing highly accurate prediction. Figure 6.16 shows results for unseen data using real faces as input.



(a)



(b)



(c)



(d)



(e)

Figure 6.16 The 1D-CNN model accurately predicts unseen data in five different categories. (a) Accuracy of prediction of unseen data predicted as Class FD1. (b) Accuracy of prediction of unseen data predicted as (Class FD2-L). (c) Accuracy of prediction of unseen data predicted as (Class FD2-R). (d) Accuracy of prediction of unseen data predicted as (Class FD3-L). (e) Accuracy of prediction of unseen data predicted as Class (FD3-R).

6.3 A 1D-ConvLSTM Model to Predict Patients at Risk of Deterioration

Long Short-Term Memory (LSTM) can handle temporal input data and achieve high accuracy of prediction. However, it fails to capture features of spatial data. Therefore, the work of (X. Shi et al., 2015a) developed the ConvLSTM model replacing the state-to-state transition operations in LSTM with convolution operations. As it involves convolution operations within the LSTM structure, it has demonstrated remarkable success in capturing and handling complex dynamic patterns within image sequences and video streams and is popular in computer vision and video analysis tasks. The ConvLSTM model expands the traditional LSTM capabilities by learning and retaining spatial dependencies in the input sequential data for tasks involving sequential data with spatial characteristics, such as video analysis, spatiotemporal modelling, and image sequence processing (R. Singh et al., 2023).

Convolution layers

These layers are responsible for performing convolution operations on input data to capture spatial patterns and relationships to extract relative features (Shi et al., 2015a; R. Singh et al., 2023).

LSTM Cells

These cells are involved in capturing temporal dependencies in the input data. Each cell includes three types of gates, the input, forget, and output gates, which are responsible for regulating the flow of information through the cell, allowing the network to retain or discard information over time (Tian et al., 2018).

By combining convolution layers and LSTM cells, the model can effectively process both spatial and temporal dependencies in the sequential data. Therefore, it is considered suitable for handling tasks like video prediction, action recognition, and FER for which both the spatial and temporal aspects of data are crucial.

This thesis proposes the ConvLSTM model for FER through frames of video stream due to its ability to capture both spatial and temporal dependencies in FEs over time. In this model, a background removal procedure was applied before the generation of the extraction vector. Then, an expressional vector was applied to detect and characterise the 5 kinds of FE of deterioration.

6.3.1 Data Pre-Processing and Design of a 1D-ConvLSTM Model

Data Pre-Processing

This section outlines the pre-processing techniques used before feeding data into the proposed ConvLSTM model. To achieve consistency across various image datasets, it is important to adjust image dimensions by resizing them to a specific size.

For consistency between image datasets, it is crucial to resize the images to a particular size. The next crucial step in an AFER pipeline is detecting and identifying faces, providing relative data by focusing on the face region of interest to classify FEs types.

To avoid processing undesirable data such as noise and irrelevant data, we applied the background removal procedure by extracting the faces from the frames of all the databases using the MediaPipe Face detection algorithm (refer to Figure 6.17). The obtained sequences of faces are aligned and then resized to 224*224 pixel resolution and converted to RGB, as required by MediaPipe and other libraries and tools, such as Matplotlib and PIL (Python Imaging Library), which expect images to be in RGB format. In the final step of pre-processing, oversampling is applied to the imbalanced training dataset.

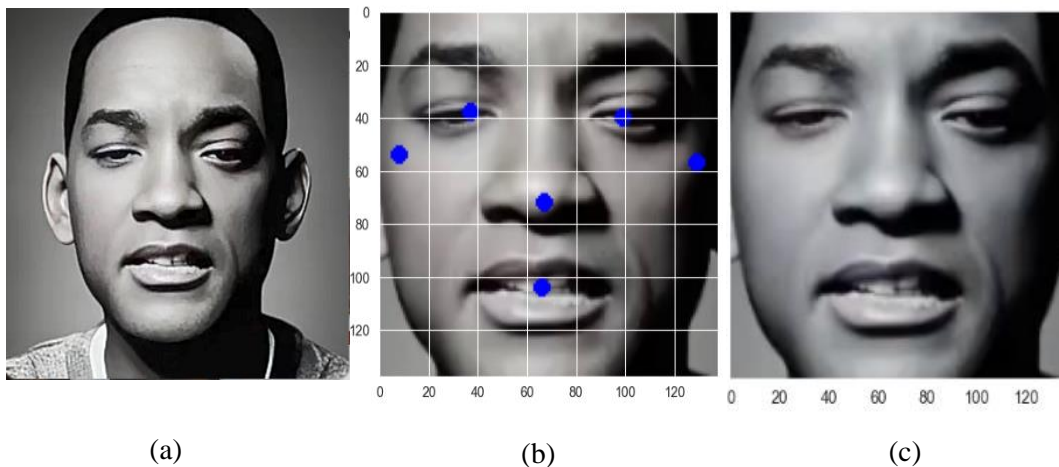


Figure 6.17 Frame of video in class FD1 before and after applying face detection. (a) Before applying Face Detection. (b) After applying Face Detection Show facial landmarks that have been used to extract face. (c) Output frame after applying face detection.

Design of a 1D-ConvLSTM

Convolution Long Short-Term Memory (ConvLSTM) is an advanced variant of LSTM that incorporates convolution operations into the network. ConvLSTM is particularly effective for spatiotemporal sequence prediction tasks where the data exhibits both spatial and temporal patterns, making it well-suited for applications like video-based FER. The proposed system stages are depicted in Figure 6.18 and details of the layers of ConvLSTM model are illustrated in Figure 6.19.

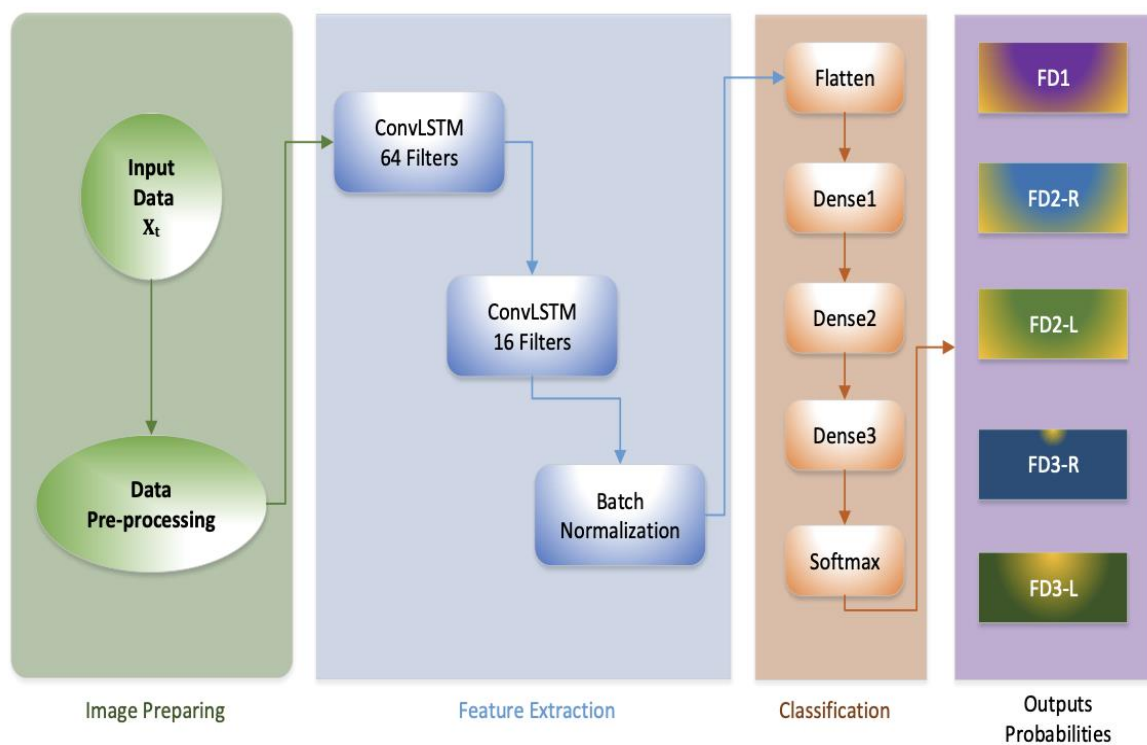


Figure 6.18 The proposed 1D-ConvLSTM model architecture.

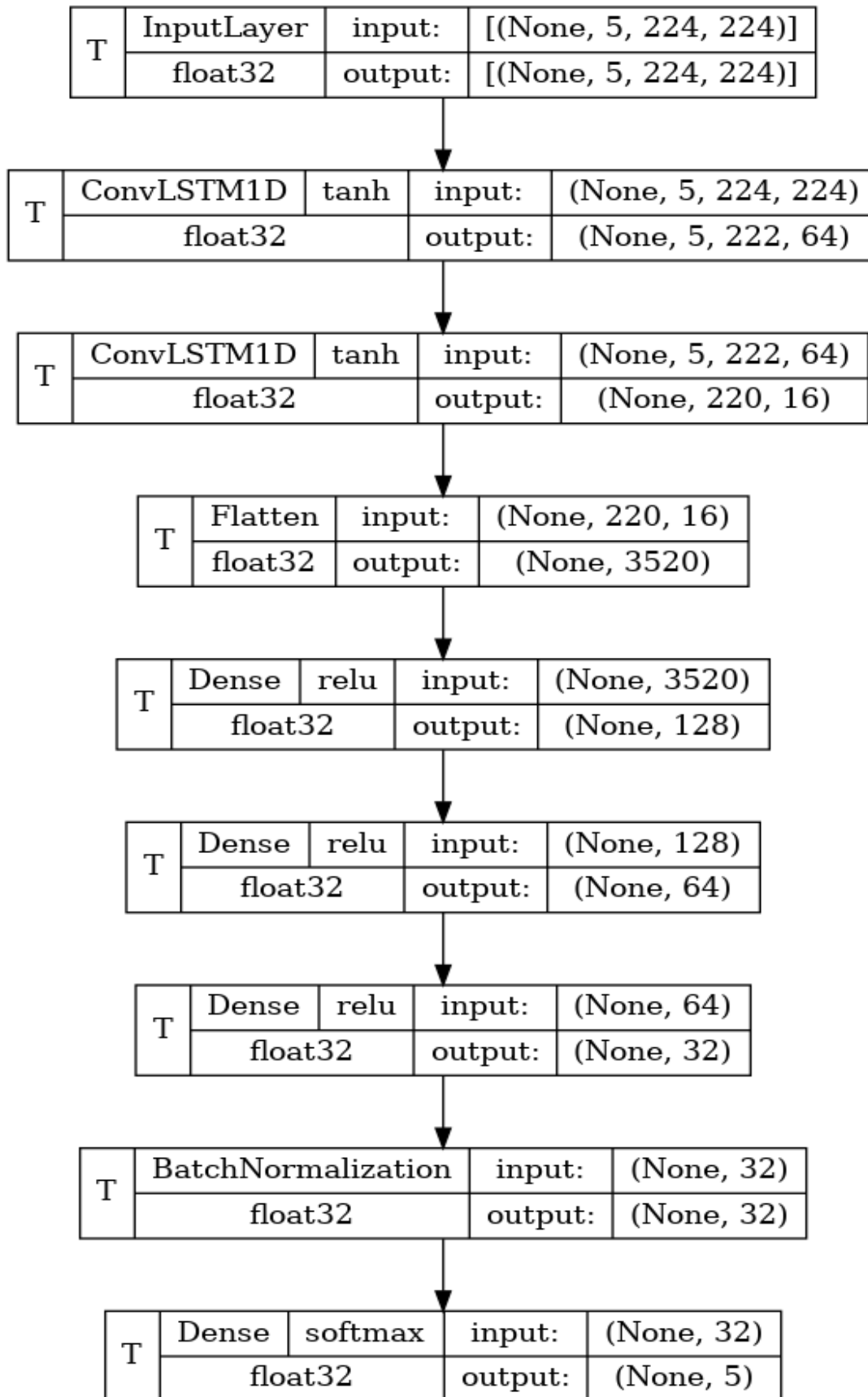


Figure 6.19 Layers of proposed 1D-ConvLSTM model.

6.3.2 Results and Evaluation of the 1D-ConvLSTM Model

Figures 6.20 and 6.21 show the accuracy and loss of the proposed model with the testing dataset.

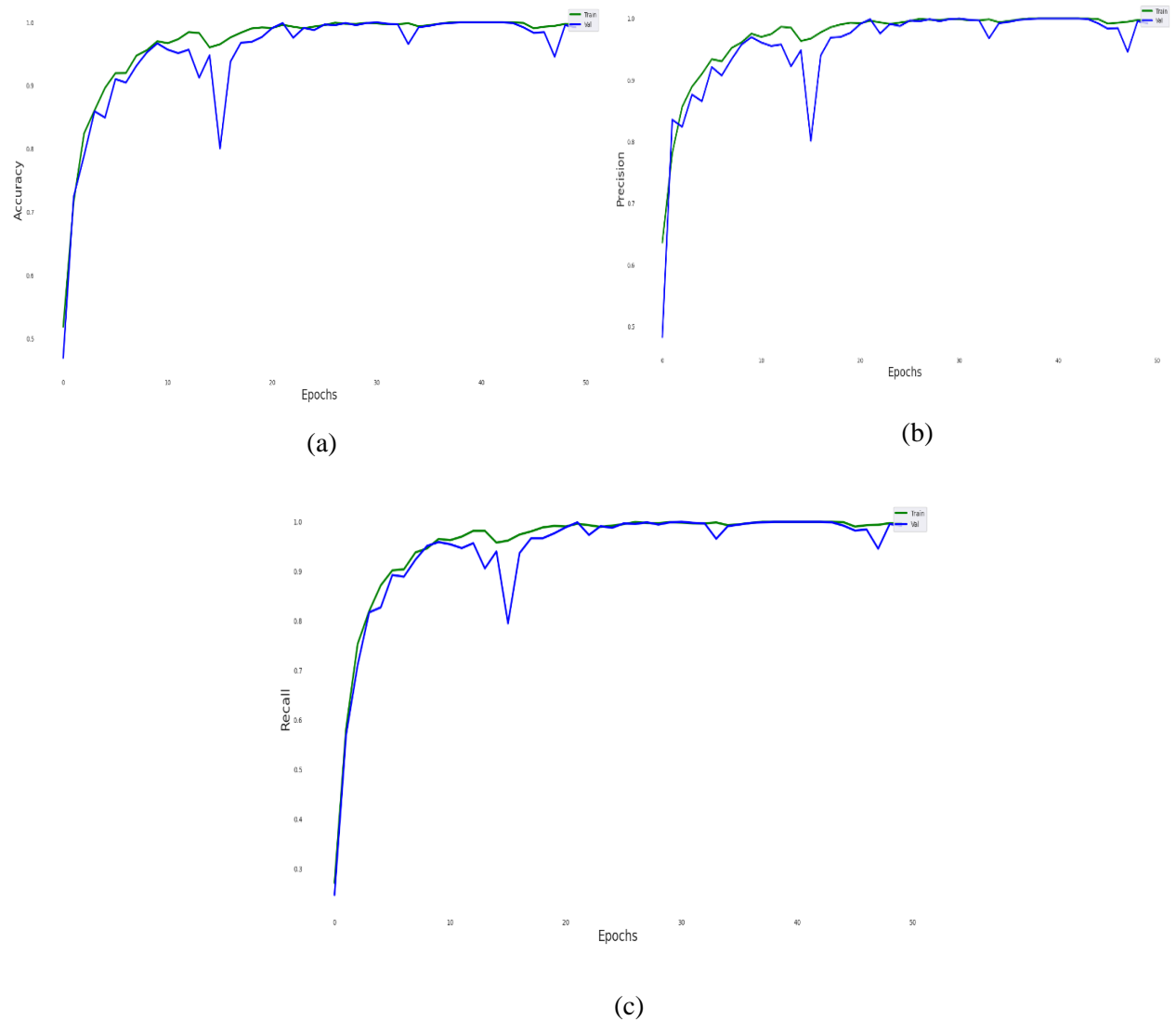


Figure 6.20 The Evaluation Metrics of 1D-ConvLSTM Model Performance. Accuracy of Proposed Model. (b) Precision of Proposed Model. (c) Recall of Proposed Model.

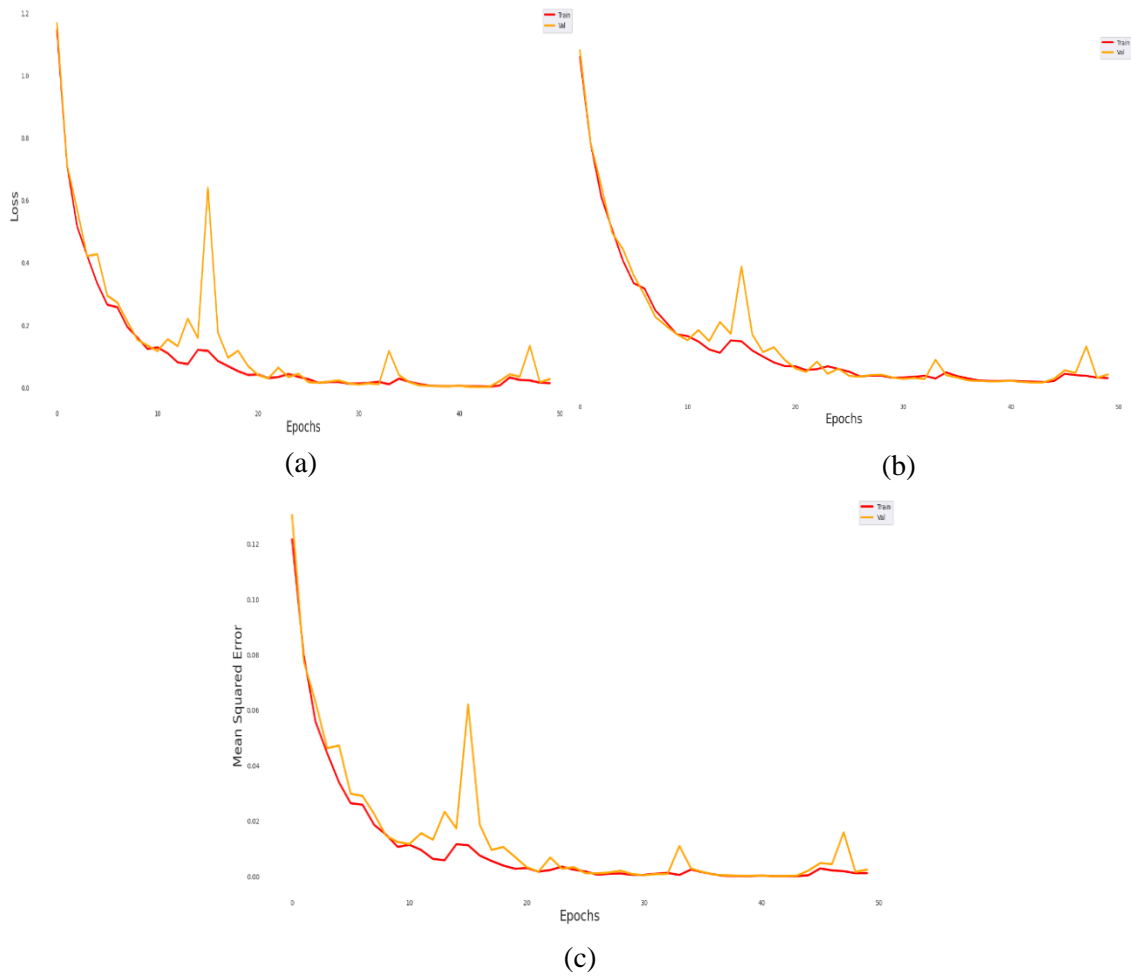


Figure 6.21 Loss, Mean Square Error, and Mean Absolute Error of 1D-ConvLSTM Model. (a) Loss of Predicted Model. (b) Mean Square Error. (c) Mean Absolute Error.

Figure 6.22 depicts the confusion matrix that summarises and visualises the performance of the proposed model across multiple classes. Matrix rows represent FEs in the real class, and columns represent FEs in the predicted class.

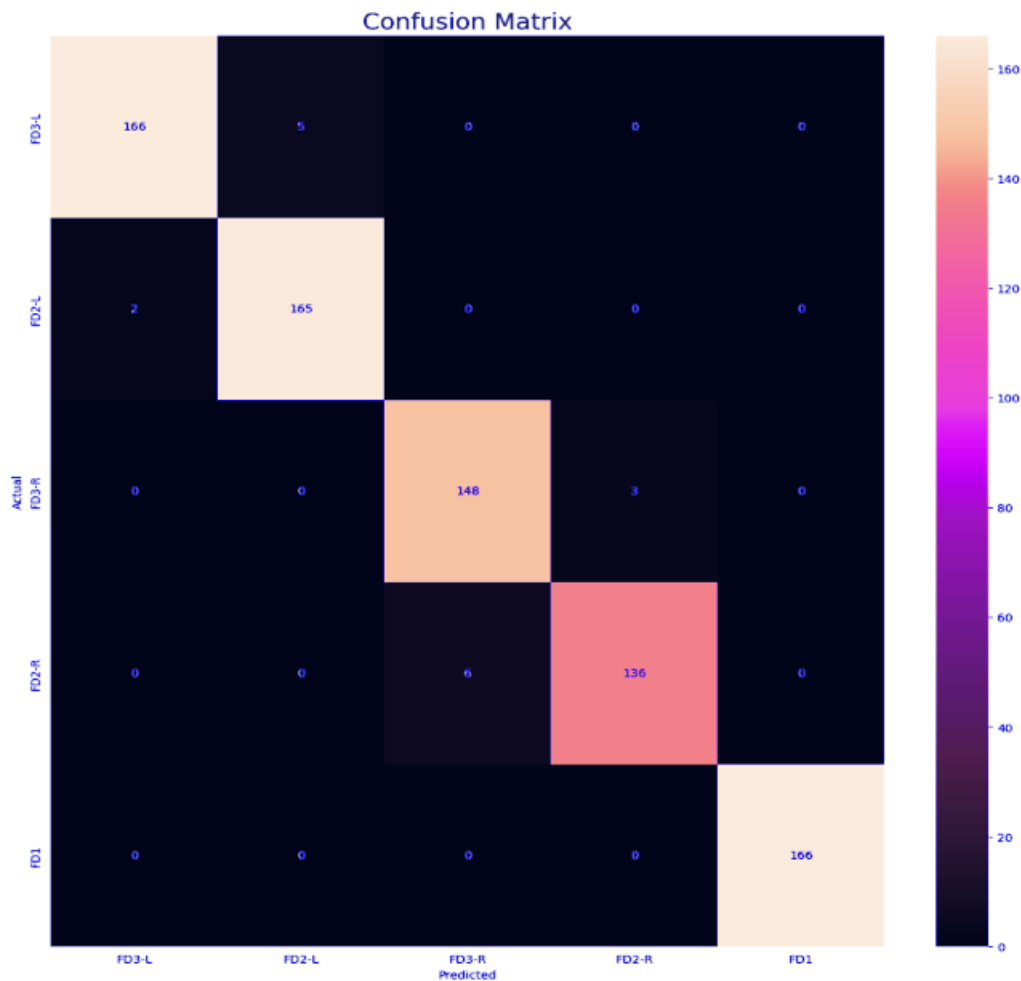


Figure 6.22 The Confusion Matrix of 1D-ConvLSTM Model.

The Confusion Matrix of the ConvLSTM model shows a high accuracy model performance. The analysis details of this matrix illustrate that TP samples of FD1, which is represented by fifth row, has 166 samples accurately identified with none FP misclassified samples. FD2-L (represent by second row) shows 165 instances are accurately classified with two samples misclassified as FD3-L. This is probably due to the features of FD2-L and FD3-L are not highly differential from each other. The same issue occurred with samples of FD2-R (fourth row), the TP is 136 instances correctly identified as FD2-R, but 6 False Positive samples are misclassified as FD3-R for the same previous reason. For FD3-L (First row), TP is 166 samples accurately classified and 5 FP instances wrongly identified as FD2-L. Lastly, FD3-R (Third row) shows 148 TP instances correctly identified as FD3-R, and 3 FP instances misclassified as FD2-R.

The key Observations form this matrix that most classes show a high rate of correct predictions (high diagonal values), indicating good model accuracy. However, still there are some

misclassification samples between FD3-R and FD2-R, and to a lesser extent between FD3-L and FD2-L. This probably occurs due to confusion between these classes, possibly due to similar features that the model fails to distinguish clearly. The high true positive rates across most classes suggest that the model has effectively learned to recognize the distinguishing characteristics of these classes. The misclassifications observed might indicate that certain features are not being adequately captured or differentiated by the model, particularly between FD3-R and FD2-R, and FD3-L and FD2-L.

Figures 6.23 and 6.24 illustrate all evaluated measurements of the proposed model.

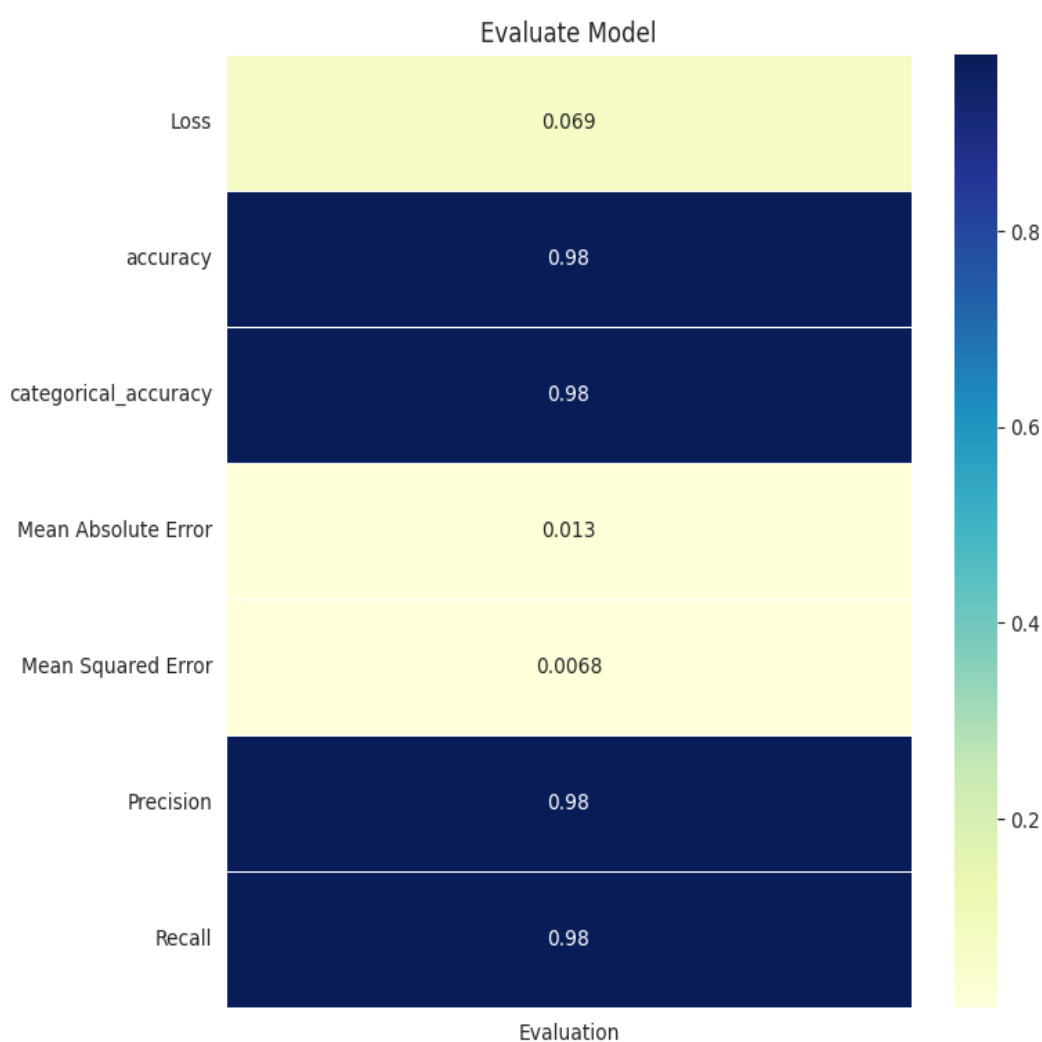


Figure 6.23 The 1D-ConvLSTM Model Evaluation Metrics.

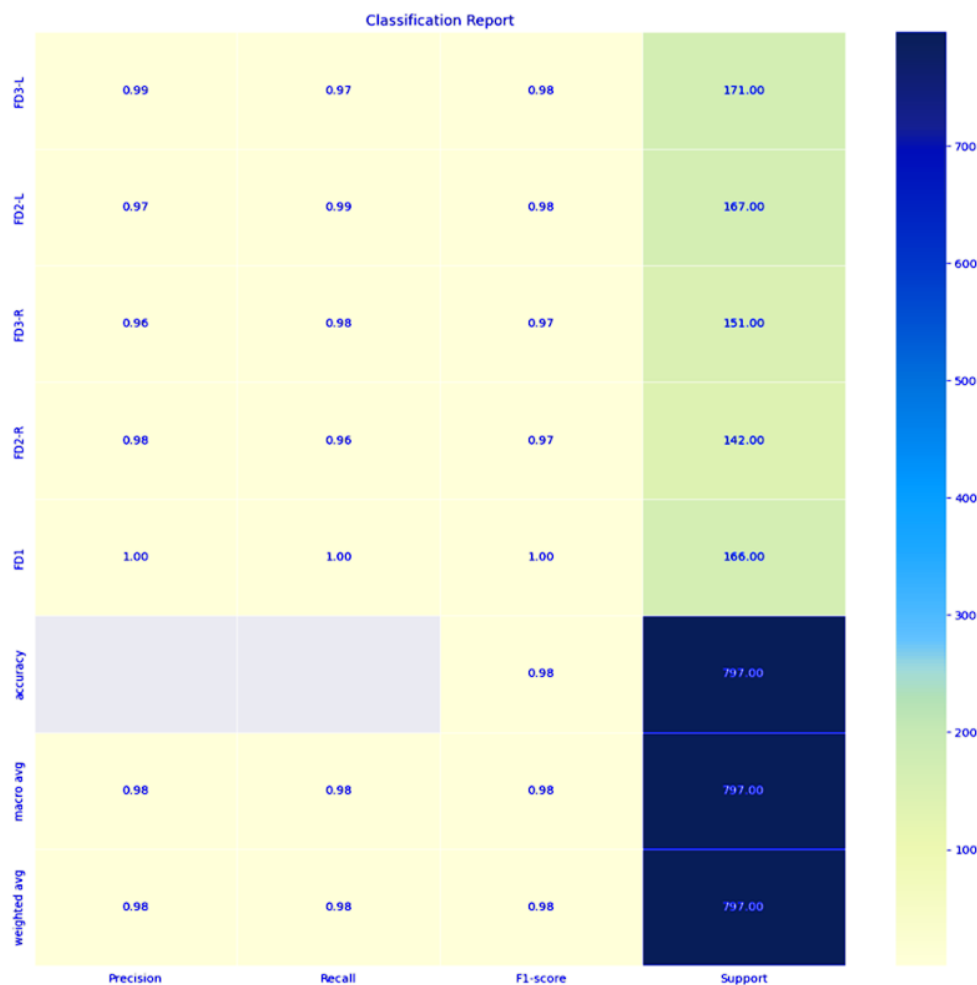


Figure 6.24 The Classification Report of 1D-ConvLSTM Model.

Figure 6.25(a) depicts the ROC for the ConvLSTM model, presenting the true positive rate (TPR) on the y-axis, the proportion of positives correctly identified, and the false positive rate (FPR) on the x-axis, the proportion of negatives incorrectly classified as positives. The ROC curve shows high classification performance for all 5 classes as their ROC curves close to the top-right corner near 1.00.

When the evaluation metrics show high accuracy, a model’s generalization ability should be confirmed using k-fold cross validation, testing the model's robustness across different subsets of the data. The model should also be validated on independent external data. In summary, although the ROC curve points to high performance, it is crucial to ensure that this reflects true predictive ability and not an artifact of dataset limitations or overfitting.

Figure 6.25(b) depicts a precision-recall (PR) curve, a key diagnostic tool used to evaluate the performance of a classification model, especially useful when dealing with imbalanced datasets. Precision is on the y-axis, measuring the ratio of correct positives to all positives, and

recall is on the x-axis, measuring the ratio of correct positives to actual positives, as detailed in Chapter 3.

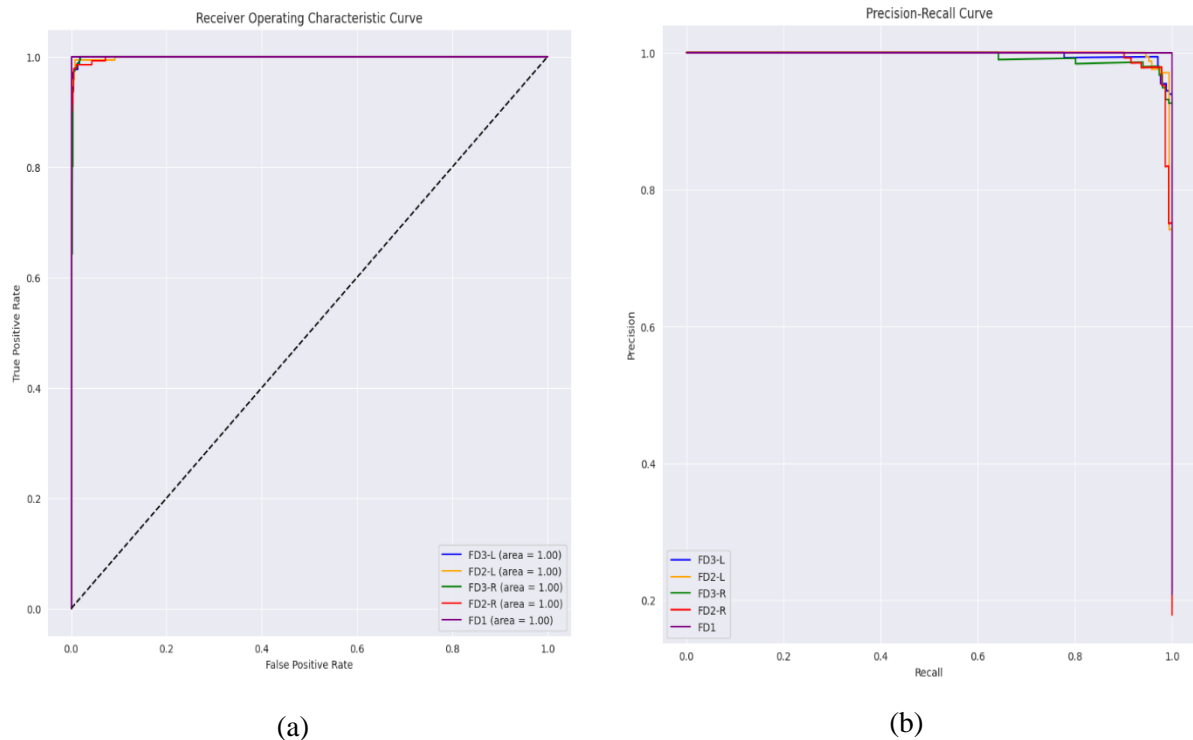


Figure 6.25 Evaluating 1D-ConvLSTM Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve. (a) ROC. (b) Precision-Recall Curve.

The PR curves show high precision and recall as the curves for all FD classes are positioned close to the top-right corner. This indicates that the model correctly classifies a high percentage of positives. The curves reflect the model performance, showing highly accurate classification across the classes as the model reaches both high recall and high precision simultaneously. In addition, the model displays minimal trade-off between recall and precision until close to the end of the recall spectrum. This suggests effective learning and classification capabilities, particularly in distinguishing between positive and negative classes while maintaining an ability to detect nearly all positives. Furthermore, the curves for different classes overlay closely, indicating consistent model performance across different categories. This uniformity and stability across various classes indicate the model's robustness and its ability to generalise across different types of data within the same task. In summary, the high precision and recall values suggest that the model is probably highly effective in practical applications as it can reliably identify positive cases with few errors.

Concerning the model's drawbacks, the sharp drop at the extreme end of the recall spectrum highlights the model's limitations when pushed to detect every possible positive. This could be critical in applications where missing a positive case has severe consequences. The threshold for classification can be optimised to balance precision and recall according to the specific application needs, especially if a small decrease in precision for a gain in recall is acceptable. The model can also be tested under different conditions and with different data subsets to ensure consistent performance that does not depend on specific characteristics of the training data.

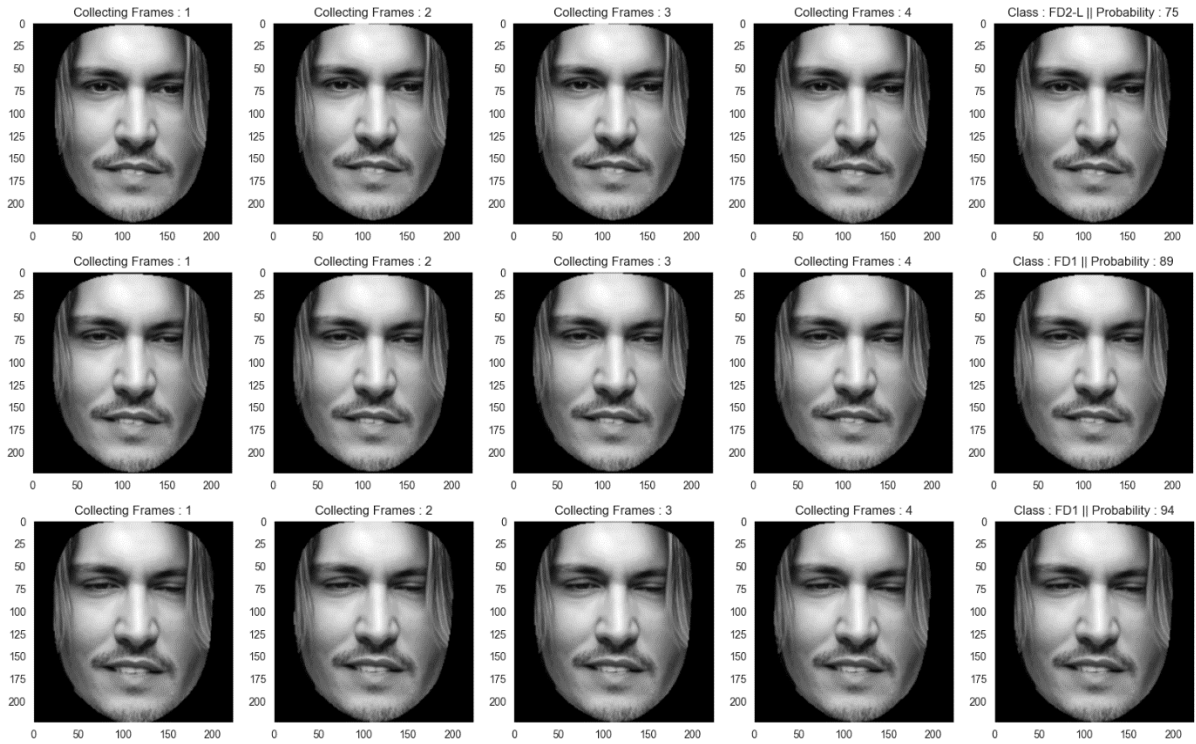
Overall, the PR curve indicates an excellent model performance, but careful consideration of application-specific requirements and potential edge cases is necessary to fully leverage its capabilities in real-world scenarios. The five evaluation measurements for the five classes under investigation are illustrated in Table 6.2.

Table 6.2 Evaluation Metrics for each Class & The Total Mean Performance of 1D-ConvLSTM.

Class Name	Facial Expression	Precision	Recall	F1 Score	Accuracy
FD1	AU (15+43+25)	100%	100%	100%	100%
FD2-R	AU (15+43+55)	98%	96%	97%	97%
FD2-L	AU (15+43+56)	97%	99%	98%	99%
FD3-R	AU (15+25+43+55)	96%	98%	97%	97%
FD3-L	AU (15+25+43+56)	99%	97%	98%	98%
	Mean	98%	98%	98%	98.2%

The above measurements provide insights into various aspects of model performance, recording 98% precision, 98% recall, 98% F1-score, and 98.2% accuracy. The model was also evaluated on an unseen dataset, showing highly accurate prediction.

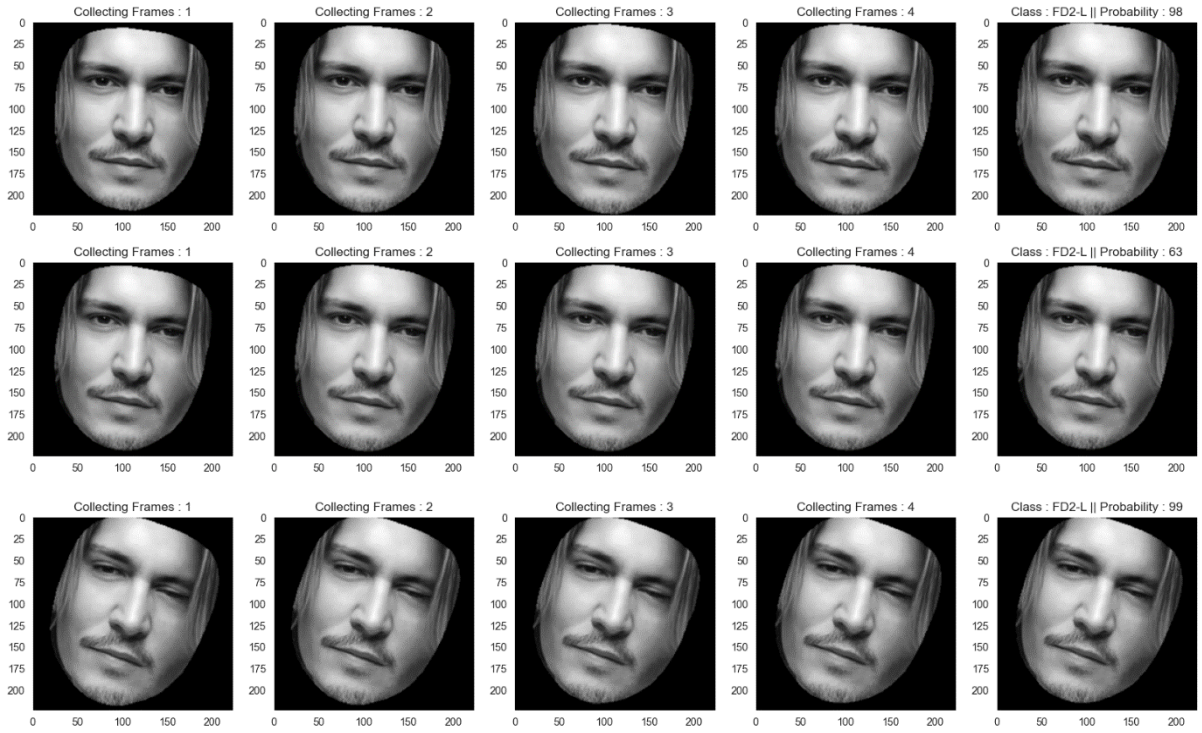
Figure 6.26 shows model's prediction for some unseen facial frames.



(a)



(b)



(c)



(d)



(e)

Figure 6.26 The model accurately predicts unseen data in five different categories.

- (a) Accuracy of prediction of unseen data predicted as Class FD1.
- (b) Accuracy of prediction of unseen data predicted as (Class FD2-L).
- (c) Accuracy of prediction of unseen data predicted as (Class FD2-R).
- (d) Accuracy of prediction of unseen data predicted as (Class FD3-L).
- (e) Accuracy of prediction of unseen data predicted as Class (FD3-R).

In conclusion, the designed model can reliably predict deteriorated facial expressions; it can reliably be used in the future in the development of health care systems where it will entirely rely on data collected from patients in critical care units. The work proposed here will fill a needs gap as, to date, there are no studies or published work that used ML in detecting deterioration through facial expressions.

In conclusion, the designed model can reliably predict deterioration FEs and can reliably be used to develop health care systems entirely reliant on data collected from patients in critical care units. This work fills a research and need-gap as, to date, there are no published studies on ML in detecting deterioration through facial expressions.

Generally, previous work relies on self-reporting. However, this approach suffers from low reliability and is unfeasible for regular use, being time consuming, impractical for monitoring patients throughout the day, and costly, requiring professional nurses and psychologists. This thesis uses automatic prediction using ML and DL based on the self-reporting work of Madrigal et al. Finally, although the overall performance of the model was remarkable, extra

information related to AUs in other face parts needs to be explored to mitigate, for example, the limitations of a face mask. The system designed in this thesis requires information from the lower part of the face.

6.4 Transformers

This section explores whether the vision transformer architecture would be suitable for automatic FER. The same procedures for data preprocessing and feature extraction used in the previous two DNN models, CNN and ConvLSTM, are applied in this approach before feeding the data to a vision transformer (ViT) model using PyTorch. Figure 6.27 depicts an overview of the data pipeline developed and tested.

Results and Discussion

Unfortunately, the generated dataset is not suitable for this transformer model, so the model fails to learn and display meaningful results. Therefore, this task is left for future investigations into other transformer models that may show better performance for FER such as Swin Transformer and emotion-attentive Transformer.

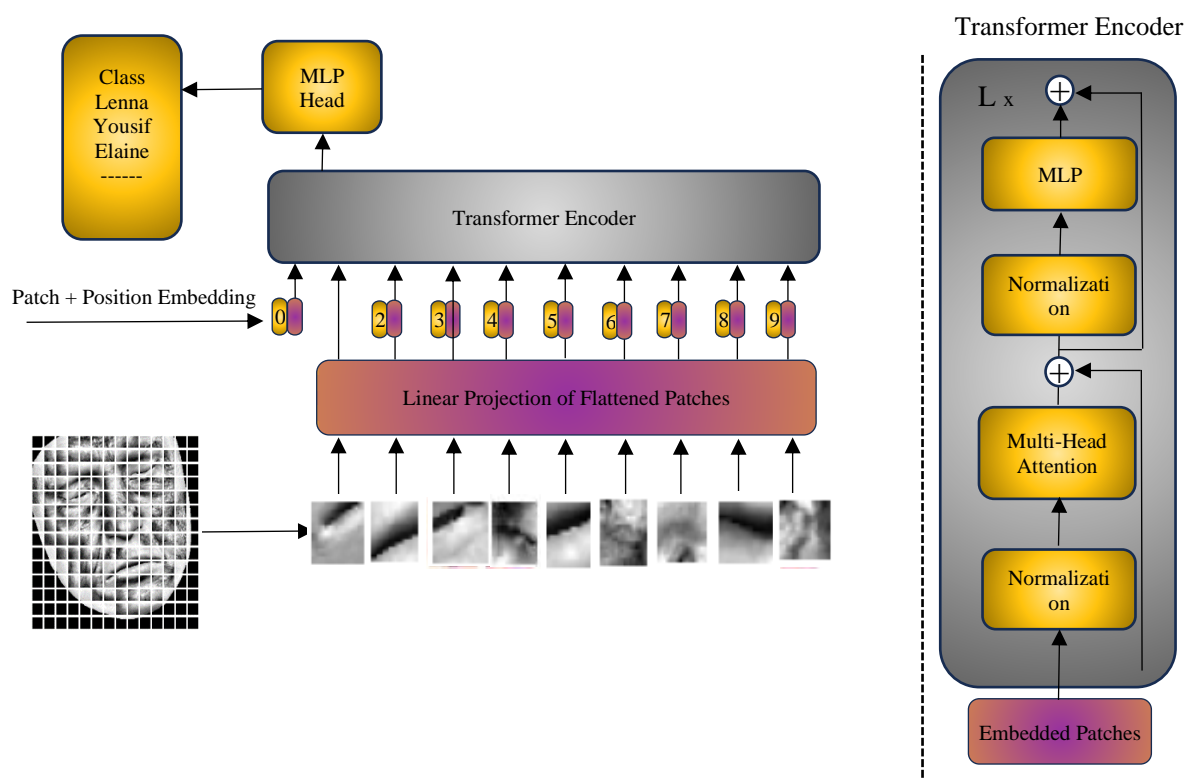


Figure 6.27 Overview system (Transformer-based FER).

Figures 6.28 and 6.29 show the accuracy and loss of the proposed model for the testing dataset, and Figures 6.30, 6.31, and 6.32 illustrate the evaluation metrics ROC, precision-recall curve, and confusion matrix.

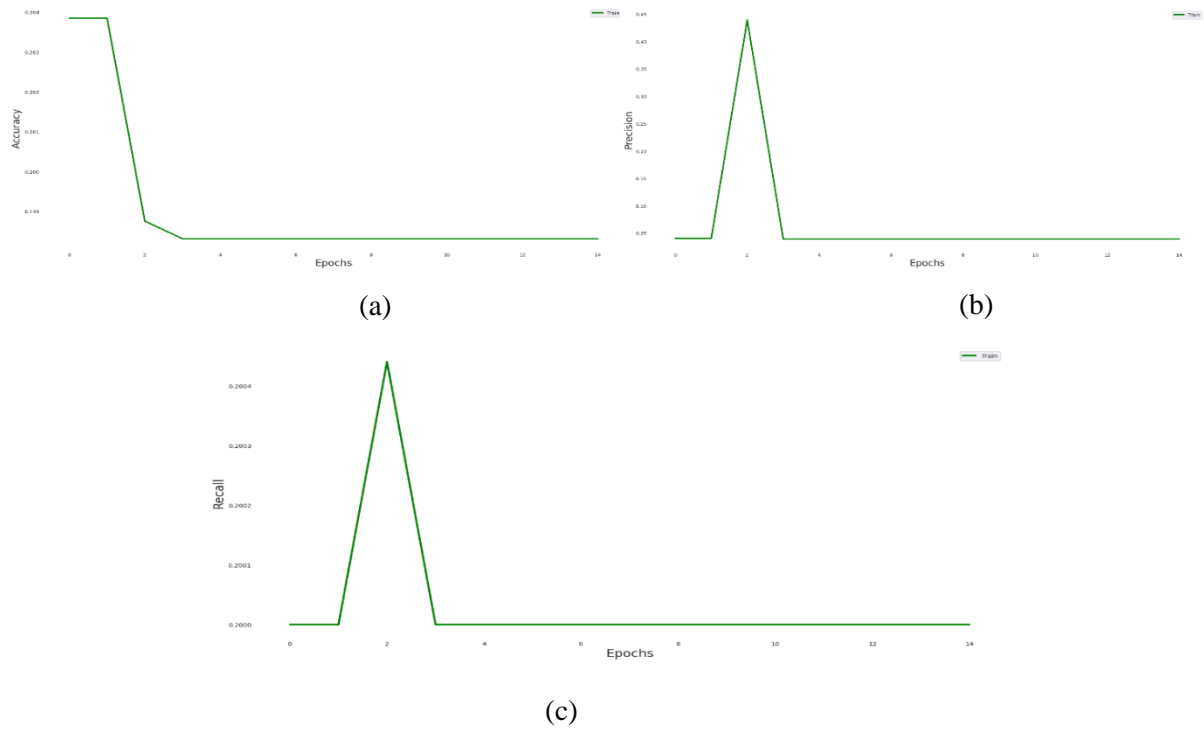


Figure 6.28 Evaluation Metrics of Model Performance. (a) Accuracy of Proposed Model. (b) Precision of Proposed Model. (c) Recall of Proposed Model.

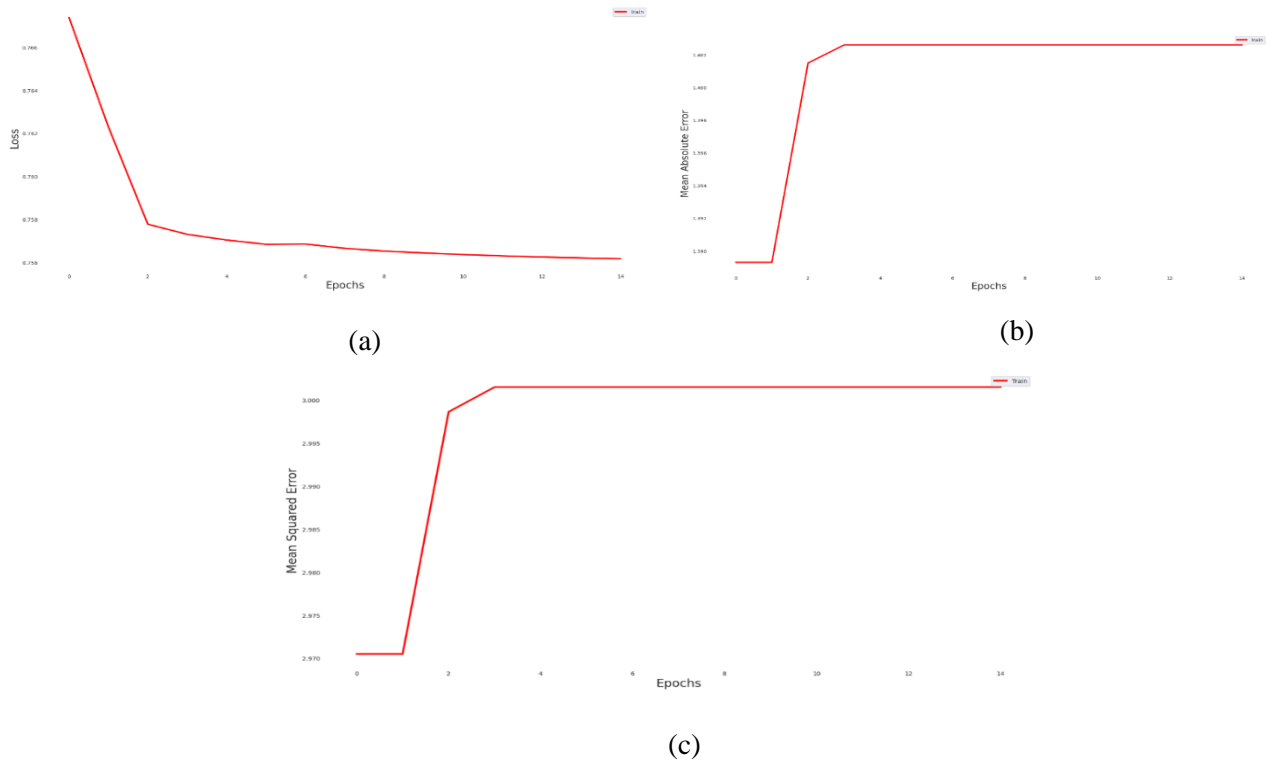


Figure 6.29 Loss, Mean Square Error, and Mean Absolute Error of ViT Model. (a) Loss. (b) Mean Square Error. (c) Mean Absolute Error.

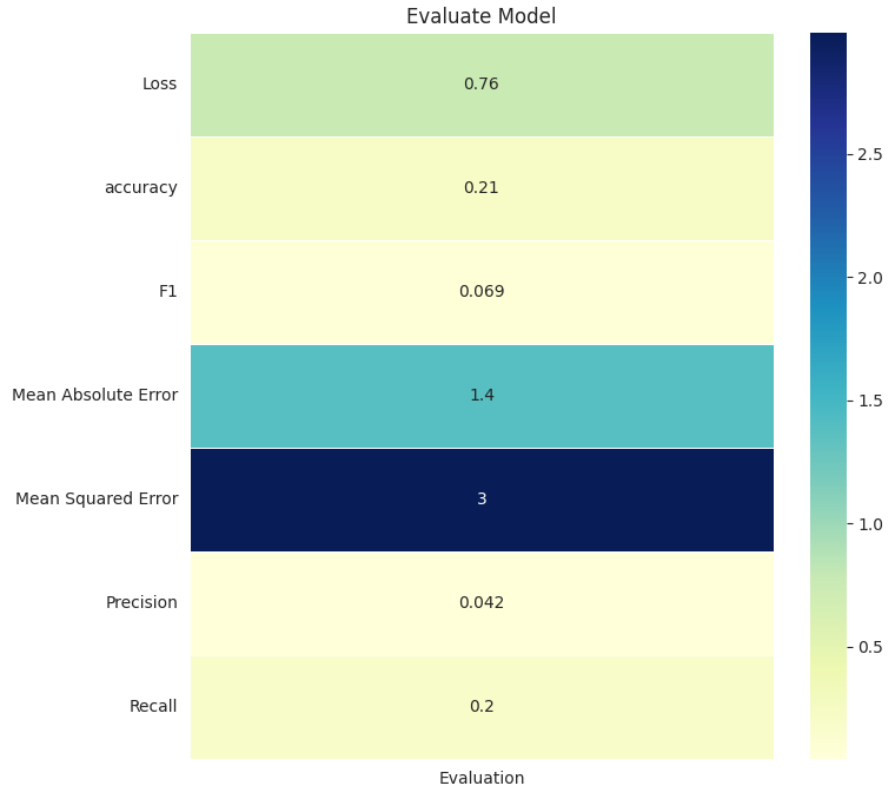


Figure 6.30 The Evaluation Metrics of ViT Model.

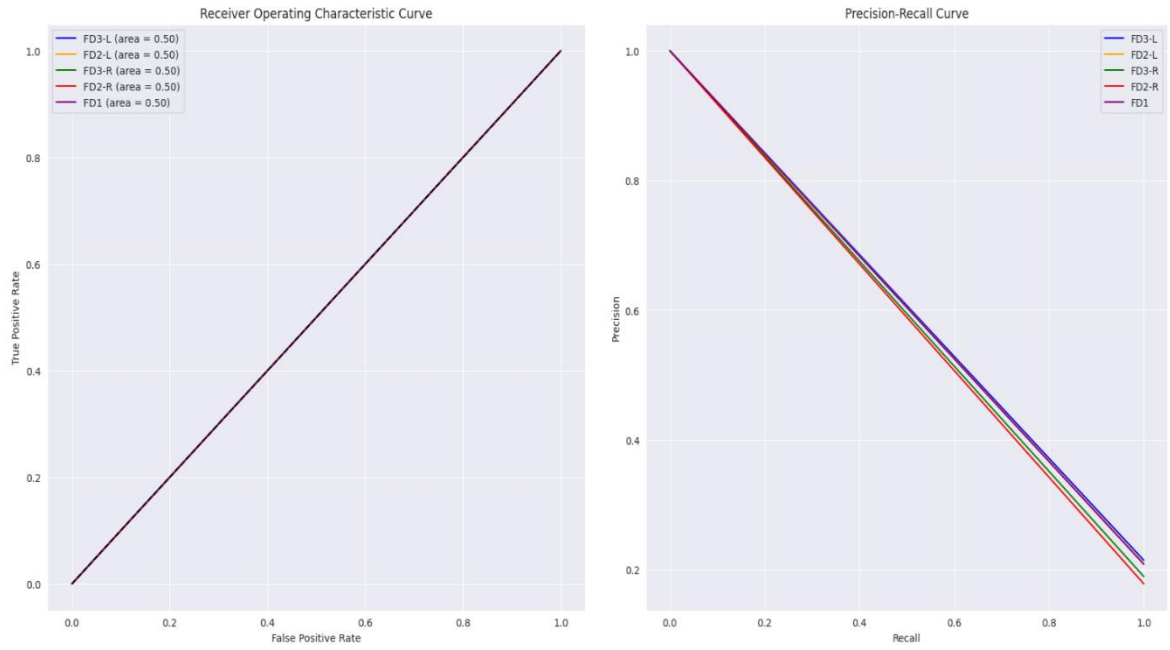


Figure 6.31 Evaluating ViT Model by Receiver Operating Characteristics Curve (ROC) & Precision-Recall Curve. (a) ROC. (b) Precision-Recall Curve.

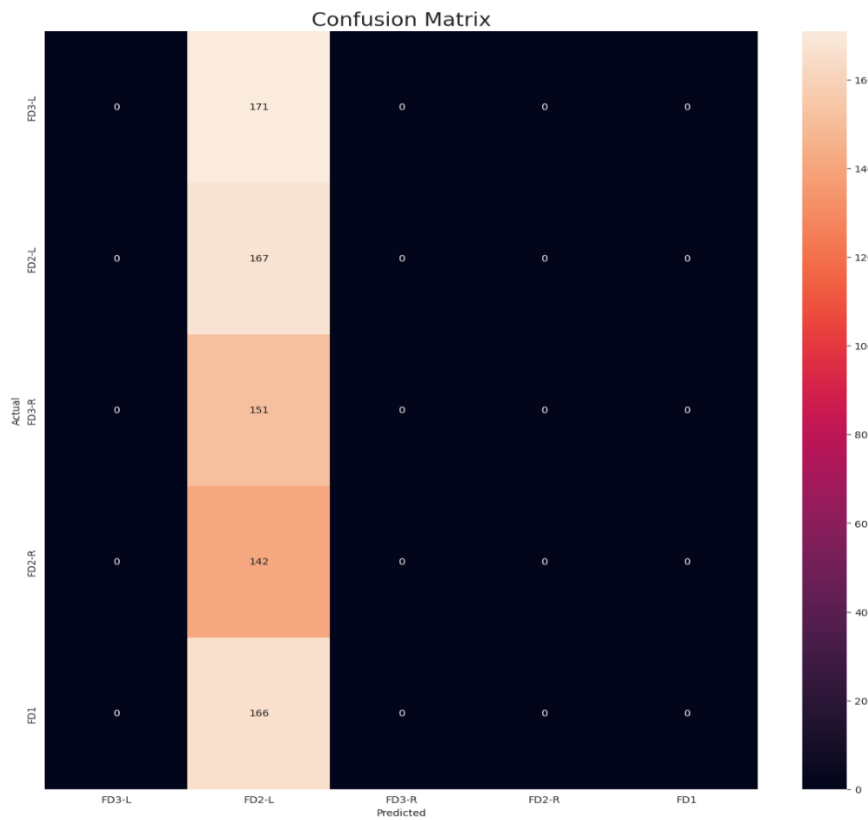


Figure 6.32 The Confusion Matrix of Vision Transformer Model.

6.5 Model Comparative Analysis

In this section, we compare the performance of machine learning (ML) and deep learning (DL) models in facial expression recognition (FER).

Machine Learning Models

For ML models, both RF and SVM models used a pre-trained facial mesh model as a feature extraction method. The accuracy of SVM was around 60%, while RF reached a prediction accuracy of 100% on the validation dataset. Comparative observations on these two classifiers are as follows.

1. SVM struggled with nuanced FEs that began after four seconds of video, especially when using other feature extraction methods that have difficulties recognising and extracting the subtle patterns.
2. RF achieved remarkable accuracy using the facial mesh model as a feature extraction method. However, RF is highly sensitive to imbalanced classes. For the first several experiments, the generated dataset was imbalanced, one of the classes was significantly larger than the others, and the accuracy did not cross 90%. Consequently, the data in the smaller classes was expanded and oversampling was employed to provide balanced data across all classes. To conclude, it is important to address this RF limitation using techniques such as random oversampling, random under-sampling, or synthetic minority oversampling Technique (SMOTE).

Deep Learning Models

The accuracy of convolutional neural networks (CNNs) is over 99%. The CNN model succeeded in capturing intricate spatial patterns, leading to better classification of subtle expressions while overfitting was mitigated using regularisation techniques like dropout and batch normalisation.

The ConvLSTM model, a combination architecture of CNN and LSTM leveraging both spatial and temporal data, achieved significant accuracy with high precision prediction around 98.8%.

In general, DL models outperform traditional ML models due to their ability to capture complex patterns without the need for feature extraction methods. However, with hybrid ML/DL models, using DL for feature extraction and ML for classification, the AFER model achieved remarkable accuracy in the same range of precision as DNN models.

In this thesis, the CNN model was pre-trained on large datasets to detect faces and facial landmarks. The extracted features were then fed to RF and SVM classifier to categorise features to their FE classes.

However, ML models remain suitable for smaller datasets where DL models may overfit. FER systems need robust generalisation for real-world conditions, like variable lighting and pose, and subtle changes in FEs. Ensemble approaches and transfer learning from pre-trained models significantly improved model robustness and accuracy.

Limitations of DNN models

Overfitting is a concern in DL models, especially with insufficient or imbalanced data. Regularisation techniques, data augmentation, and cross-validation helped address these issues. In addition, complex DNN models with large datasets require high computational resources. Some FEs remain challenging to consistently classify due to micro FEs and overlaps in features between classes.

The performance of the models SVM, RF, 1D-CNN, ConvLSTM, and Transformers were compared. First, the model is evaluated using the generated database PRD-FE applying the k-fold cross-validation process.

Genetic algorithm (GA) did not apply as already the 1D-CNN and ConvLSTM have achieved the state of the art in predicting the 5 FEs.

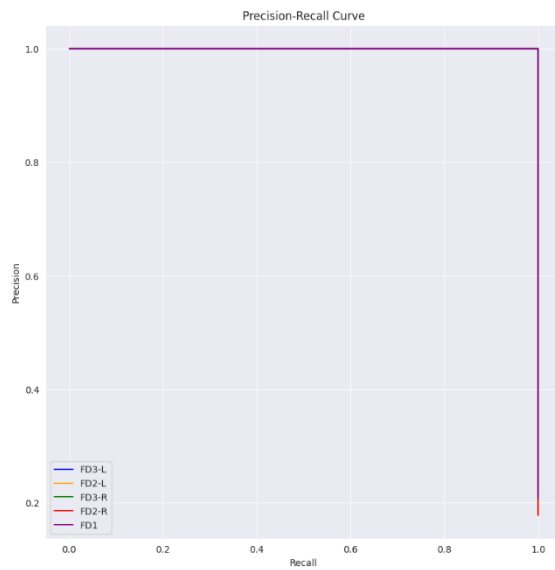
Performance Evaluation Criteria

Table 6.3 shows the accuracy (the proportion of correctly classified samples) of the 5 models investigated in this thesis.

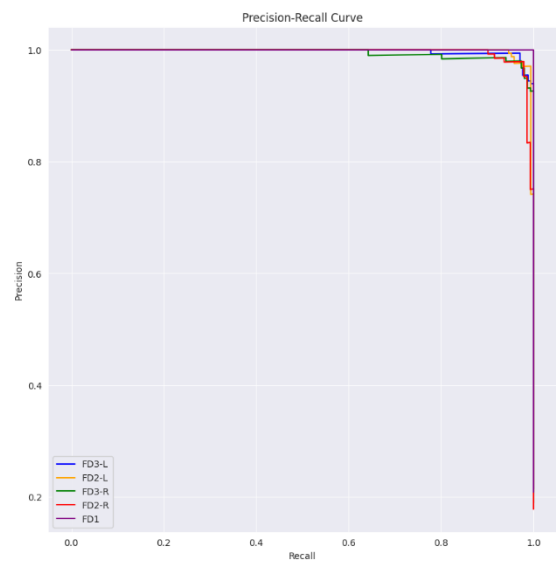
Table 6.3 Accuracy of the models using generated database PRD-FE

Accuracy				
1D-CNN	1D-ConvLSTM	Transformer	Random Forest	Support Vector Machine
99.74%	99.89%	20.95%	100%	60%

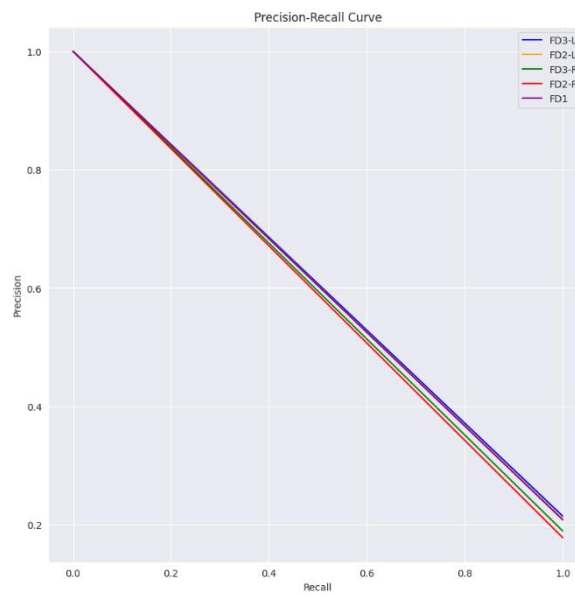
Figure 6.33 shows precision and recall for the three DNN models, indicating the model's ability to categorise specific FE.



(a) Precision-Recall Curve of CNN model.



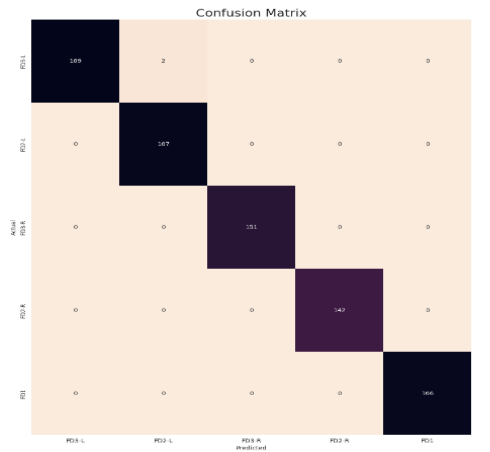
(b) Precision-Recall Curve of ConvLSTM model.



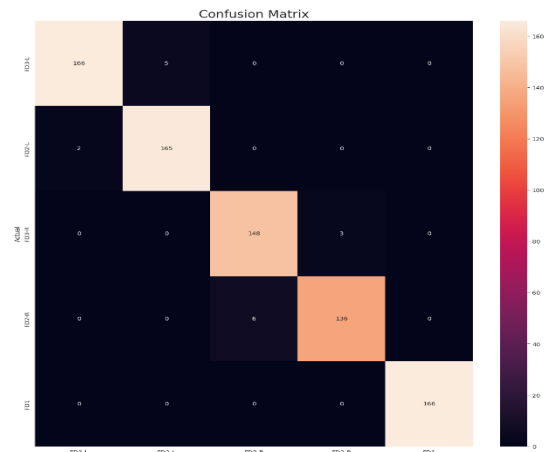
(c) Precision-Recall Curve of Transformer model.

Figure 6.33 The Precision-Recall Curve for Each DNN model.

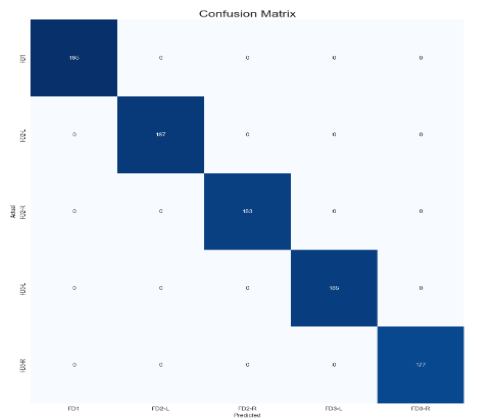
Figure 6.34 illustrates Confusion Matrix of each ML and DL models to analyse and detect which FE are often misclassified.



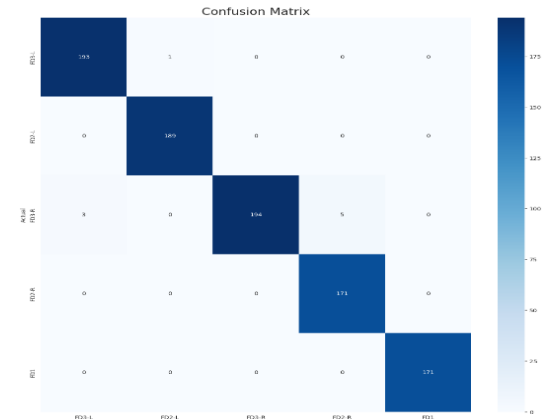
(a) Confusion Matrix of CNN model.



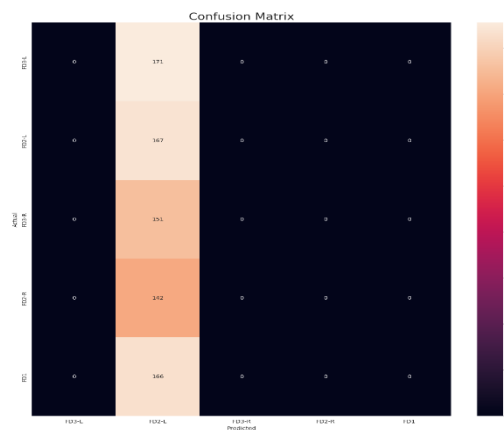
(b) Confusion Matrix of ConvLSTM model.



(c) Confusion Matrix of RF model.



(d) Confusion Matrix of SVM model.



(e) Confusion Matrix of Transformer.

Figure 6.34 Confusion Matrix for each proposed model.

Table 6.4 shows the evaluation metrics of the DL model.

Table 6.4 The overall error rate of the designed model of predicting FEs of patients at risk of Deterioration.

Model Type	MSE	MAE
(a)1D-Conv	0.0008	0.0040
(b)1D-ConvLSTM	0.0001	0.0006
(c)Transformers	3	1.4

Figure 6.35 depicts a receiver operating characteristic (ROC) curve for three different DNN models. ROC is a graphical plot commonly used to evaluate and illustrate the performance of a classification model at various threshold settings. The ROC curve plots the true positive rate (TPR, also known as sensitivity) against the false positive rate (FPR, also known as 1-specificity) at different threshold levels.

For Figure 6.35a and 6.35b show both ROC curves of 1D-CNN and 1D-ConvLSTM for all classes are located at the top-right corner, indicating both high precision and high recall with values close to 1.0 (the models correctly classified almost all positive examples without incorrectly labelling negative examples as positives). High precision and high recall indicate that the models are both accurate and comprehensive in their positive classifications.

For Figure 6.35c, the curves for all FD classes FD1, FD2-L, FD2-R, FD3-L, and FD3-R have an area under the curve (AUC) of 0.50, which indicates that the model's ability to discriminate between the positive and negative classes is like random guessing. This is represented by the diagonal line from (0,0) to (1,1), which serves as a baseline to indicate no discriminative power. The model represented here does not provide any meaningful separation between the classes and is apparently unable to learn any relevant patterns from the dataset.

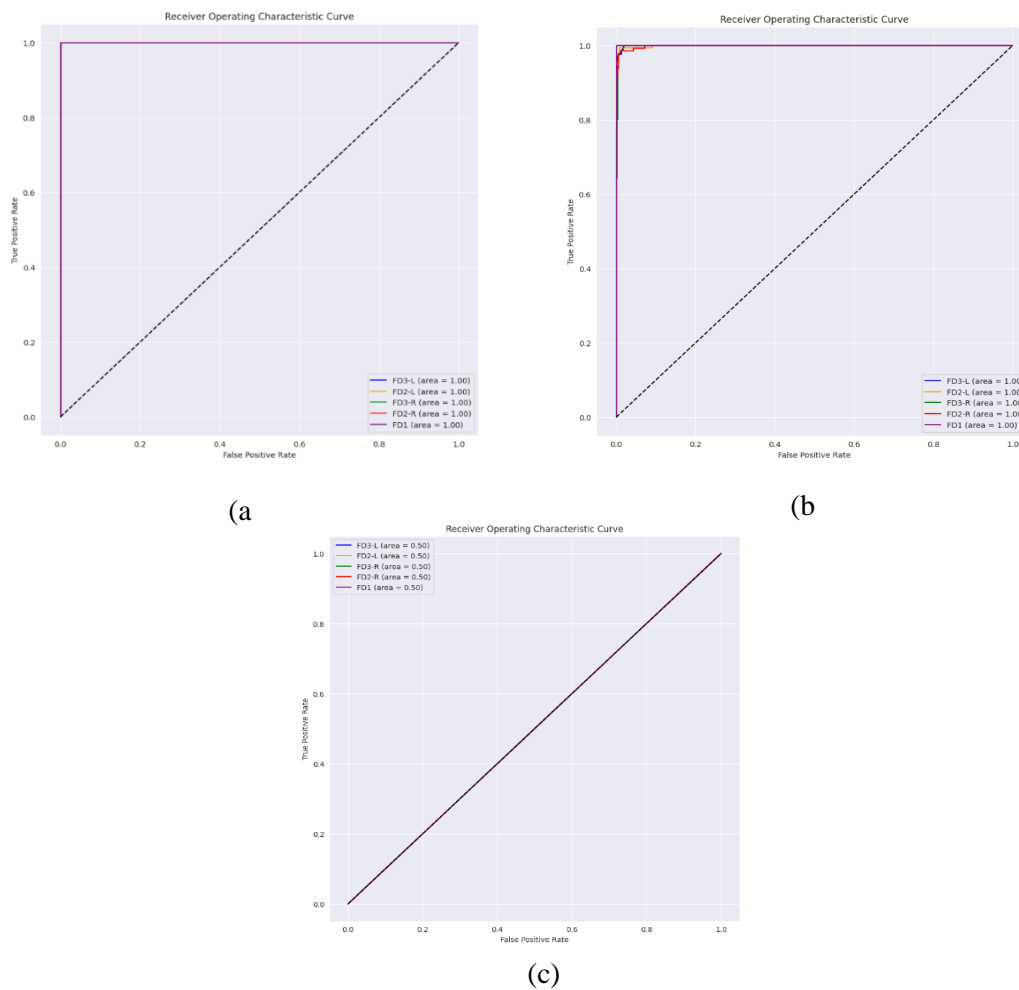


Figure 6.35 Comparison between evaluated Models by Receiver Operating Characteristics Curve (ROC). (a) ROC of 1D-Conv. (b) ROC of 1D-ConvLSTM. (c) ROC of Transformers.

6.6 Conclusions

In conclusion, the proposed DNN models, 1D-CNN and 1D-ConvLSTM, can reliably predict the FEs of patients at risk of deterioration and can be applied in a clinical test to develop a fully validated healthcare system reliant only on real data collected from patients. Real patient facial data was emulated using a variety of public celebrity faces. If only real patient data were to be used in a clinical trial, it would be important to consider whether the model complexity is appropriate for the training data available and, if necessary, to adjust to prevent overfitting. To the best of our knowledge, to date, no published study has designed a ML or DL model to detect FEs of patients at risk of deterioration in hospital wards or critical care units. Although the overall performance of the models was remarkable, real dataset samples should be collected in the future to better understand the model limitations.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

Our faces hold and show valuable clues about human emotions and their intentions. FER has been intensively studied for the last few decades in computer vision due to its importance to improving communication with individuals and generate empathetic responses. Early detection of signs of patients' deterioration from facial expressions is a challenging task for healthcare professionals. Therefore, this thesis has concentrated on proposing suitable methods employing deep learning algorithms as a solution for identifying signs of patients' deterioration through their FEs. With recent technologies and advancement in computer vision, pattern recognition, and machine learning, it is possible to detect and characterize FEs through images and video streams with high accuracy using DNN models such as CNN and LSTM networks.

The main objective of this research was to design a framework for automatic FER to predict FEs of patients at risk of deterioration. The proposed system used a generated database called PRD-FE comprising five different combination sets of AUs, (FD1, FD2-R, FD2-L, FD3-R and FD3-L) representing FEs of deterioration risk. Employing the DNN models for face detection and ML models for facial landmark detection significantly improves model performance.

We anticipate applying this model to the development of health care systems within critical care units. Therefore, future work will concentrate on collecting real-world data to further validate the proposed models and present them as integrated systems with other medical assessment systems to enhance the chances of human survival.

A literature review (Chapter 2) identified and critically evaluated studies that are related to the scope of this project. Based on this review, statistical methods for identifying deteriorated facial expressions have been addressed along with their limitations. In addition, the main automatic approaches for designing models that can predict patients under pain, autism, etc., through facial expressions, and their limitations have been identified. Based on the literature it has been concluded that there is a need to design an automatic facial expression recognition that can trigger warning signs as indicators of patients who are under risk of deterioration.

Introducing automatic facial expression recognition model can improve health care systems and human survival by reducing the reliance on self-reporting methods, which can be done by full use of advanced technology as has been demonstrated by this thesis. In addition, the obstacle of occlusion by face mask has been addressed. It has also been highlighted that there is a need to collect more information by psychologists related to appearance of FE in the upper part of patient's face to make AFER systems more robust in recognizing the set of AUs as a deterioration status and distinguish them from other FEs such as drowsiness expression.

The literature review has led this thesis to investigate deep learning methods, to save cost and time from psychologists, improve human survival by timely intervention, and recognize subtle changes of deteriorating facial expressions at early stages. Results from this investigation have shown that automatic facial expression recognition can reliably improve human survival as an integrated part of health care system in intensive care units. It would replace the self-reporting system, which sometimes depends on intuitive decisions and naked eye perceptions making it difficult to recognize subtle changes at the early stages of deterioration.

Three deep learning models have been evaluated and tested along with using various pre-processing approaches to examine their effectiveness to predict collapsing facial expressions at risk of deterioration. For SVM, RF and the evaluation matrixes, the free Python machine learning library scikit-learn was used (<https://scikit-learn.org/stable/index.html>). The CNN and ConvLSTM models were implemented using Keras API, which is a Python open-source high-level API for TensorFlow (https://www.tensorflow.org/api_docs/python/tf/keras). For testing the Vision Transformer, PyTorch libraries were used for building and training the model.

One of the proposed DNN models combines a 1D-CNN with LSTM enabling the extraction of higher feature patterns using 1D-CNN, while also extracting the sequential correlations in the input sequences of time series with LSTM. In addition, using face detection and face oval selection techniques enabled the model to achieve high accurate prediction with average 98% accuracy, prior to the appearance of facial expressions at risk of deterioration.

The 1D-CNN model recorded better performance and accuracy compared with other DNN models. Using 1D-CNN model in conjunction with face detection and face oval MediaPipe techniques as a preprocessing method results in attaining the significant accurate prediction scored of 99.8%. The two previous models have achieved state-of-the-art in predicting FEs of deterioration.

Overall, this thesis work on deep learning represents significant advances to the modelling of facial expressions recognition to improve chances of human survival in critical care settings. The models proposed in this thesis can be used to develop healthcare applications by using state-of-the-art deep learning methods enabling reliable detecting of deterioration trends with timely and targeted intervention.

7.2 Findings and the Impact of the Project

It is important in healthcare system to develop real time alert system that can automatically recognise early markers of deterioration, which triggers notifications to healthcare providers when a patient's data indicates potential deterioration, allowing early intervention and improved patient outcomes. The literature review also indicated that there are main limitations to statistical approaches in recognizing deteriorated facial expressions; first at the early stages of deterioration, using self-reporting lacks the ability to recognise subtle changes of facial expressions due to observing through naked eyes. This certainly can miss valuable information resulting in losing the benefits of timely intervention. Second, it is also expected that the self-reporting may fail in accurately predicting deterioration through FEs as assessment is often based on intuitive decisions. Therefore, the literature review highlighted the importance of employing ML and DL as an AFER method to achieve accurate prediction. In addition, the findings of this thesis found that some DNN models can be accurately used in predicting subtle changes at the early stage of deterioration. Furthermore, combining DL models with preprocessing techniques helps to capture minute changes during micro facial expressions.

Consequently, this thesis has developed an automatic FER model that detects FEs of patients at risk of deterioration. At the first stage of the project, a synthetic database has been generated based on the work of Madrigal and other contributors (Madrigal-Garcia et al., 2018) which identified that, when the patients are under deterioration, the facial muscles start collapsing and show specific facial expressions as an indicator of deterioration. This thesis developed realistic 3D models of human faces that can accurately mimic FEs associated with patient deterioration with various features such as age, skin tone, facial landmarks, ethnicity, etc.

By implementing techniques such as rotation, scaling, different intensity of AUs, and expression blending, this thesis was able to expand the diversity of FEs in the dataset. After that, key facial landmarks and features were annotated to ensure accurate representation of expressions, as this is crucial to both the generation process and the training stage of deep learning models. Then, using First order Motion Model technology to transfer the subtle facial

movements to images of real human faces, ensuring that the synthetic expressions are as realistic as possible. The generated dataset is created through unlimited conditions such as environment simulation created with variations in lighting, camera angles, and background to create a more robust dataset that can generalize well to real-world scenarios. In addition, this thesis has ensured that the dataset includes avatars representing diverse demographics (age, gender, ethnicity) to avoid bias and improve model generalization. After generating dataset, various preprocessing methods have been exploited to improve model performance such as face detection techniques as they help in improving model robustness and performance by avoiding undesirable data that may affect the learning process. The next stage was designing customized Convolutional Neural Networks (CNNs) and Long-Short Term Memory that can successfully capture spatial and temporal features of FEs.

The designed and demonstrated solutions open the door to the possibility of automatic prediction of patients at risk of deterioration, which will in turn enable improving human survival, timely and targeted intervention by healthcare professionals. Future work can also include models of understanding and recognizing patterns of deteriorated facial expressions despite face masks, by including relevant information from other areas of the face.

7.3 Recommendation for Future Work

While this work shows significant results in modelling facial expression recognition at risk of deterioration, there are limitations to the models as proposed. Although the designed solution was able to predict deterioration of patients through facial expressions, there remains a lack of distinguishing between drowsiness expression, for instance, that can be confused with expressions indicating risk of deterioration. When the lower part of the face is occluded with a face mask, essential features and information will be missed. In this case, the upper part may display similar features for both drowsiness and deterioration expressions making them undistinguishable as highlighted in Chapter 4. In other words, further work is required to collect additional data by psychologists from other parts of the face that are out of range of face mask to mitigate this obstacle. In addition, there is a need to evaluate and validate the design system on real patients' data to verify whether there is any further work needed to maintain current key advantages regarding robustness and high accuracy of prediction. Furthermore, the investigation of further facial AUs from other parts of the face especially the upper part, that are related to risk of deterioration. This could lead to an improved model that can detect deterioration even for patients wearing masks.

References

- Abdoli, S., Cardinal, P., & Lameiras Koerich, A. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, 136, 252–263. <https://doi.org/10.1016/j.eswa.2019.06.040>
- Abo-Tabik, M. A., Abdulmunem, M., & Abo-Tabik, S. (2021). USING DEEP LEARNING PREDICTIONS OF SMOKERS' BEHAVIOUR TO DEVELOP A SMART SMOKING-CESSATION APP. Manchester Metropolitan University.
- Adegun, I. P., & Vadapalli, H. B. (2020). Facial micro-expression recognition: A machine learning approach. *Scientific African*, 8. <https://doi.org/10.1016/j.sciaf.2020.e00465>
- Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. In *Electronics (Switzerland)* (Vol. 9, Issue 8, pp. 1–53). MDPI AG. <https://doi.org/10.3390/electronics9081188>
- Agarap, A. F. (2018). Deep Learning using Rectified Linear Units (ReLU). <http://arxiv.org/abs/1803.08375>.
- Akeh, L. J., Chandra, R. K., Loo, W., & Sutoyo, R. (2022, September). Modelling Emotions Recognition from Facial Expression using Vision Transformer with IMED Dataset. In *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 254–257). IEEE.
- Al Tae, E. J., & Jasim, Q. M. (2020). Blurred Facial Expression Recognition System by Using Convolution Neural Network. *Webology*, 17(2), 804–816. <https://doi.org/10.14704/WEB/V17I2/WEB17068>.
- Alasad, J., & Ahmad, M. (2005). Communication with critically ill patients. *Journal of Advanced Nursing*, 50(4), 356–362. <https://doi.org/10.1111/j.1365-2648.2005.03400.x>.
- Albert Mehrabian. (1972). Nonverbal communication. *Behaviour Research and Therapy*, 11(4), 669.
- Allaert, B., Bilasco, M., & Djeraba, C. (2019). Micro and macro facial expression recognition using advanced Local Motion Patterns. *IEEE Transactions on Affective Computing*, 13(1), 147–158. <https://doi.org/10.1109/TAFFC.2019.2949559i>.
- Al-Nuimi, A. M., & Mohammed, G. J. (2021). Face Direction Estimation based on Mediapipe Landmarks. *7th International Conference on Contemporary Information Technology and Mathematics, ICCITM 2021*, 185–190. <https://doi.org/10.1109/ICCITM53167.2021.9677878>.
- Al-Tekreeti, Z., Moreno-Cuesta, J., Madrigal Garcia, M. I., & Rodrigues, M. A. (2024, August). AI-

- Based Visual Early Warning System. In *Informatics* (Vol. 11, No. 3, p. 59). MDPI.
- Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., & Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4), 1-23.
- Arul Vinayakam Rajasimman, M., Manoharan, R. K., Subramani, N., Aridoss, M., & Galety, M. G. (2023). Robust Facial Expression Recognition Using an Evolutionary Algorithm with a Deep Learning Model. *Applied Sciences (Switzerland)*, 13(1). <https://doi.org/10.3390/app13010468>
- Assari, M. A., & Rahmati, M. (2011). Driver drowsiness detection using face expression recognition. 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 337–341.
- Barroso, E., Santos, G., & Proença, H. (2013). Facial expressions: Discriminability of facial regions and relationship to biometrics recognition. 2013 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), 77–80.
- Bartneck, C., & Lyons, M. J. (2007). HCI and the face: Towards an art of the soluble. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4550 LNCS(PART 1), 20–29. https://doi.org/10.1007/978-3-540-73105-4_3
- Baru, C., Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (2019). Application of multi-channel 3D-cube successive convolution network for convective storm nowcasting. 2019 IEEE International Conference on Big Data (Big Data), 1705–1710.
- Beaudry, O., Roy-Charland, A., Perron, M., Cormier, I., & Tapp, R. (2014). Featural processing in recognition of emotional facial expressions. *Cognition and Emotion*, 28(3), 416–432. <https://doi.org/10.1080/02699931.2013.833500>
- Bennett, C. C., & Šabanović, S. (2014). Deriving Minimal Features for Human-Like Facial Expressions in Robotic Faces. *International Journal of Social Robotics*, 6(3), 367–381. <https://doi.org/10.1007/s12369-014-0237-z>
- Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia*, 50(12), 2830–2838. <https://doi.org/10.1016/j.neuropsychologia.2012.08.010>
- Bombardi, D., Schmid Mast, M., Canadas, E., & Bachmann, M. (2015). Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. In *Frontiers in Psychology* (Vol. 6). Frontiers Media S.A. <https://doi.org/10.3389/fpsyg.2015.00869>
- Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning LSTM model for

- electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7). <https://doi.org/10.3390/en11071636>
- Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2020). Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting. *Energies*, 13(2). <https://doi.org/10.3390/en13020391>
- Buisine, S., Courgeon, M., Charles, A., Clavel, C., Martin, J. C., Tan, N., & Grynszpan, O. (2014). The Role of Body Postures in the Recognition of Emotions in Contextually Rich Scenarios. *International Journal of Human-Computer Interaction*, 30(1), 52–62. <https://doi.org/10.1080/10447318.2013.802200>
- Calder, A. J., Keane, J., Young, A. W., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2), 527–551. <https://doi.org/10.1037/0096-1523.26.2.527>
- Carbon, C. C. (2020). Wearing Face Masks Strongly Confuses Counterparts in Reading Emotions. *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.566886>
- Carcagnì, P., Del Coco, M., Leo, M., & Distantè, C. (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1). <https://doi.org/10.1186/s40064-015-1427-3>.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; pp. 213–229.
- Cascella, M., Schiavo, D., Cuomo, A., Ottaiano, A., Perri, F., Patrone, R., Migliarelli, S., Bignami, E. G., Vittori, A., & Cutugno, F. (2023). Artificial Intelligence for Automatic Pain Assessment: Research Methods and Perspectives. In *Pain Research and Management (Vol. 2023)*. Hindawi Limited. <https://doi.org/10.1155/2023/6018736>.
- Chandran, P., Zoss, G., Gotardo, P., & Bradley, D. (2024, May). Infinite 3D Landmarks: Improving Continuous 2D Facial Landmark Detection. In *Computer Graphics Forum* (p. e15126).
- Chang, J.-Y., & Chen, J.-L. (1999). A Facial Expression Recognition System Using Neural Networks. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, 3511–3516.
- Chen, C. H., Lee, I. J., & Lin, L. Y. (2015). Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Research in Developmental Disabilities*, 36, 396–403. <https://doi.org/10.1016/j.ridd.2014.10.015>
- Chen, J., Lv, Y., Xu, R., & Xu, C. (2019). Automatic social signal analysis: Facial expression

- recognition using difference convolution neural network. *Journal of Parallel and Distributed Computing*, 131, 97–102. <https://doi.org/10.1016/j.jpdc.2019.04.017>
- Chen, T., Yin, H., Yuan, X., Gu, Y., Ren, F., & Sun, X. (2021). Emotion recognition based on fusion of long short-term memory networks and SVMs. *Digital Signal Processing: A Review Journal*, 117. <https://doi.org/10.1016/j.dsp.2021.103153>
- Chen, X., Yang, X., Wang, M., & Zou, J. (2017). Convolution Neural Network for Automatic Facial Expression Recognition. *IEEE International Conference on Applied System Innovation*, 814–817.
- Chen, Y., & Joo, J. (2021). Understanding and Mitigating Annotation Bias in Facial Expression Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14980–14991.
- Chen, Y., Yang, Z., & Wang, J. (2015). Eyebrow emotional expression recognition using surface EMG signals. *Neurocomputing*, 168, 871–879. <https://doi.org/10.1016/j.neucom.2015.05.037>
- Choi, H.-C., & Oh, S.-Y. (2006). Realtime Facial Expression Recognition using Active Appearance Model and Multilayer Perceptron. *2006 SICE-ICASE International Joint Conference. IEEE*, 5924–5927.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). *Emotion_recognition_in_human-computer_interaction*. *IEEE Signal Processing Magazine*, 18, 32–80.
- Culjak, I., Abram, D., Pribanic, T., Dzapo, H., & Cifrek, M. (2012). A brief introduction to OpenCV. *Proceedings of the 35th International Convention MIPRO*, 1725–1730.
- Darwish, A., Ezzat, D., & Hassanien, A. E. (2020). An optimized model based on convolutional neural networks and orthogonal learning particle swarm optimization algorithm for plant diseases diagnosis. *Swarm and Evolutionary Computation*, 52. <https://doi.org/10.1016/j.swevo.2019.100616>
- Dash, M., Liu, H., & Yao, J. (1997, November). Dimensionality reduction of unsupervised data. In *Proceedings ninth IEEE international conference on tools with artificial intelligence* (pp. 532–539). IEEE.
- David, L., Samuel, M. P., Eduardo, Z. C., & Garcia-Bermejo, J. G. (2014). Animation of Expressions in a Mechatronic Head. *ROBOT2013: First Iberian Robotics Conference: Advances in Robotics*. Springer., 2, 15–26. <http://www.springer.com/series/11156>
- Davison, A., Merghani, W., Lansley, C., Ng, C. C., & Yap, M. H. (2018). Objective micro-facial movement detection using FACS-Based regions and baseline evaluation. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 642–

649. <https://doi.org/10.1109/FG.2018.00101>.

- Devlin, J., Chang, M-W., Lee, K., Toutanova, K., (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Dewmini, A., Hirshan, R., & Kumara, W. (2021). A Study of Facial Emotion Recognition Techniques to Examine Micro-Expressions. Faculty of Technology, South Eastern University of Sri Lanka, University.
- Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE., 2106–2112.
- Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (p. 29).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. <http://arxiv.org/abs/2010.11929>
- Duan, C., & Luo, S. (2022). Design of Pedestrian Detection System based on OpenCV. Proceedings - 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing, AIAM 2022, 256–259. <https://doi.org/10.1109/AIAM57466.2022.00055>.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4), 169-200.
- Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage* (revised edition). WW Norton & Company.
- Ekman, P., & Friesen, W. V. (1969). *Nonverbal Leakage and Clues to Deception* (Vol. 32).
- Ekman, P., & Friesen, W. V. (1971). CONSTANTS ACROSS CULTURES IN THE FACE AND EMOTION '. In *Journal of Personality and Social Psychology* (Vol. 71, Issue 2).
- Ekman, P., & Friesen, W. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Palo Alto*, 3, 1–5. <https://doi.org/10.1037/t27734-000>
- Ekman, P., & Friesen, W. V. (1976). Measuring-Facial-Movement. *Environmental Psychology and Nonverbal Behavior*, 1, 56–75.
- Ekman, P., Friesen, W. V. ;, & Hager, J. (2002). *Facial action coding system Research Nexus*. Network Research Information, Salt Lake City, UT, 1.
- Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar?. *American psychologist*, 46(9), 913.
- Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, 37(2), 505–515. <https://doi.org/10.1148/rg.2017160130>
- Falavigna, G. (2022). *Deep Learning for Beginners*. Moncalieri: CNR-IRCrES (Itinerari per l'alta.

- Fehrenbach, M. J., & Herring, S. W. (2015). *Illustrated Anatomy of the Head and Neck*. Elsevier Health Sciences.
- Fraser, A. S. (1957). SIMULATION OF GENETIC SYSTEMS BY AUTOMATIC DIGITAL COMPUTERS. *Australian Journal of Biological Sciences*, 10(4), 484–491.
- Fu, X., Zhang, C., Peng, X., Jian, L., & Liu, Z. (2019). Towards end-to-end pulsed eddy current classification and regression with CNN. 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE., 1–5.
- Galati, D., Scherer, K. R., & Ricci-Bitti, P. E. (1997). Voluntary Facial Expression of Emotion: Comparing Congenitally Blind With Normally Sighted Encoders. In *Journal of Personality and Social Psychology* (Vol. 73, Issue 6).
- Gilbert, M., Demarchi, S., & Urdapilleta, I. (2021). FACSHuman, a software program for creating experimental material by modeling 3D facial expressions. *Behavior Research Methods*. Springer., 53(5), 2252–2272. <https://doi.org/10.3758/s13428-021-01559-9>/Published
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1), 142–158. <https://doi.org/10.1109/TPAMI.2015.2437384>
- Gogić, I., Manhart, M., Pandžić, I. S., & Ahlberg, J. (2020). Fast facial expression recognition using local binary features and shallow neural networks. *Visual Computer*, 36(1), 97–112. <https://doi.org/10.1007/s00371-018-1585-8>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27. <http://www.github.com/goodfeli/adversarial>
- Gori, M., Schiatti, L., & Amadeo, M. B. (2021). Masking Emotions: Face Masks Impair How We Read Emotions. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.669432>
- Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. In *Vision Research* (Vol. 41). www.elsevier.com/locate/visres
- Gosselin, P., Kirouac, G., & Dore, F. Y. (1995). Components and Recognition of Facial Expression in the Communication of Emotion by Actors.
- Greche, L., Akil, M., Kachouri, R., & Es-sbai, N. (2020). A new pipeline for the recognition of universal expressions of multiple faces in a video sequence. *Journal of Real-Time Image Processing*, 17(5), 1389–1402. <https://doi.org/10.1007/s11554-019-00896-5>
- Greche, L., & Es-Sbai, N. (2016). Automatic system for facial expression recognition based histogram of oriented gradient and normalized cross correlation. 2016 International Conference on Information Technology for Organizations Development (IT4OD), 1–5.

- Gren, L., & Lindberg, D. (2024). The State of Live Facial Puppetry in Online Entertainment.
- Gross, J., Cuesta, J., & Crawford, S. (2013). The face of illness: Analysing facial expressions in critical illness in conjunction with the facial action coding system (FACS). *Intensive Care Medicine*, 39(S265).
- Gross, J., Cuesta, J., Crawford, S., Devaney, M., & Madrigal-Garcia, M. (2013). The face of illness Analysing facial expressions in critical illness in conjunction with the facial action coding system . *Intensive Care Medicine*. SPRINGER 233 SPRING ST, NEW YORK, NY 10013 USA, 39, S265–S265.
- Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Image and Vision Computing*, 28(5), 807–813. <https://doi.org/10.1016/j.imavis.2009.08.002>.
- Gülbetekin, E., Fidancı, A., Altun, E., Er, M. N., & Gürcan, E. (2023). Effects of mask use and other-race on face perception, emotion recognition, and social distancing during the COVID-19 pandemic. *Asian Journal of Social Psychology*, 26(4), 445–460. <https://doi.org/10.1111/ajsp.12570>.
- Gunes, H., & Hung, H. (2016). Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block. *Image and Vision Computing*, 55, 6–8. <https://doi.org/10.1016/j.imavis.2016.03.013>.
- Guo, X., Zhang, Y., Lu, S., & Lu, Z. (2023). Facial expression recognition: a review. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-023-15982-x>.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.
- Hachisuka, S., Kimura, T., Ishida, K., Nakatani, H., & Ozaki, N. (2010). Drowsiness detection using facial expression features (No. 2010-01-0466). *SAE Technical Paper*.
- Haggard, E. A. ;, & Isaacs, K. S. ; (1966). Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of Research in Psychotherapy* (pp. 154–165).
- Hardas, B. M., & Pokle, S. B. (2017). Optimization of Peak to Average Power Reduction in OFDM. *Journal of Communications Technology and Electronics*, 62(12), 1388–1395. <https://doi.org/10.1134/S1064226917140017>.
- Hasani, B., & Mahoor, M. H. (2017). Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 30–40.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- Helmud, E., Fitriyani, F., & Romadiana, P. (2024). Classification Comparison Performance of

- Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, 13(1), 92-97.
- Herr, K., Coyne, P. J., Ely, E., Gélinas, C., & Manworren, R. C. B. (2019). Pain Assessment in the Patient Unable to Self-Report: Clinical Practice Recommendations in Support of the ASPMN 2019 Position Statement. *Pain Management Nursing*, 20(5), 404–417. <https://doi.org/10.1016/j.pmn.2019.07.005>.
- Hickson, S., Kwatra, V., Dufour, N., Sud, A., & Essa, I. (2019). Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, 1626–1635. <https://doi.org/10.1109/WACV.2019.00178>.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. <http://arxiv.org/abs/1207.0580>.
- Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. *Acm Sigart Bulletin*, 53, 15–15.
- Huang, C. J., Shen, Y., Chen, Y. H., & Chen, H. C. (2021). A novel hybrid deep neural network model for short-term electricity price forecasting. *International Journal of Energy Research*, 45(2), 2511–2532. <https://doi.org/10.1002/er.5945>.
- Huang, X., Wang, S.-J., Zhao, G., & Pietikäinen, M. (2015). Facial Micro-Expression Recognition using Spatiotemporal Local Binary Pattern with Integral Projection. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 1–9.
- Huang, Z., Wang, S., & Yu, K. (2018). Angular softmax for short-duration text-independent speaker verification. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September*, 3623–3627. <https://doi.org/10.21437/Interspeech.2018-1545>.
- Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Machine Learning Automated Machine Learning Methods, Systems, Challenges. <http://www.springer.com/series/15602>.
- Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. <http://arxiv.org/abs/1502.03167>.
- Iyer, A., Das, S. S., Teotia, R., Maheshwari, S., & Sharma, R. R. (2023). CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. *Multimedia Tools and Applications*, 82(4), 4883–4896. <https://doi.org/10.1007/s11042-022-12310-7>.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

- Jaiswal, S., & Valstar, M. (2016). Deep Learning the Dynamic Appearance and Shape of Facial Action Units. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE., 1–8.
- Jaswanth, K. S., & David, D. S. (2020, July). A novel based 3D facial expression detection using recurrent neural network. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-6). IEEE.
- Kalal, Z., Mikolajczyk, K., & Matas, J. (2010, September). Face-tld: Tracking-learning-detection applied to faces. In 2010 IEEE International Conference on Image Processing (pp. 3789-3792). IEEE.
- Jeong, M., & Ko, B. C. (2018). Driver's facial expression recognition in real-time for safe driving. *Sensors (Switzerland)*, 18(12). <https://doi.org/10.3390/s18124270>.
- Jogin, M., Mohana, Madhulika, M. S., Divya, G. D., Meghana, R. K., & Apoorva, S. (2018). Feature extraction using convolution neural networks (CNN) and deep learning. 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2018 - Proceedings, 2319–2323. <https://doi.org/10.1109/RTEICT42901.2018.9012507>.
- Jones, D., Mitchell, I., Hillman, K., & Story, D. (2013). Defining clinical deterioration. *Resuscitation*, 84(8), 1029–1034. <https://doi.org/10.1016/j.resuscitation.2013.01.013>.
- Joseph, A., & Geetha, P. (2020). Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. *Visual Computer*, 36(3), 529–539. <https://doi.org/10.1007/s00371-019-01628-3>.
- Kalal, Z., Mikolajczyk, K., & Matas, J. (2010, September). Face-tld: Tracking-learning-detection applied to faces. In 2010 IEEE International Conference on Image Processing (pp. 3789-3792). IEEE.
- Kartynnik, Y., Ablavatski, A., Grishchenko, I., & Grundmann, M. (2019). Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. <http://arxiv.org/abs/1907.06724>.
- Kazemi, V., & Josephine, S. (2014). One Millisecond Face Alignment with an Ensemble of Regression Trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874.
- Khanum, H., & Pramod, H. B. (2022). Smart Presentation Control by Hand Gestures Using computer vision and Google's Mediapipe. *International Research Journal of Engineering and Technology (IRJET)*, 2657.
- Kim, T. Y., & Cho, S. B. (2018). Web traffic anomaly detection using C-LSTM neural networks. *Expert Systems with Applications*, 106, 66–76. <https://doi.org/10.1016/j.eswa.2018.04.004>
- Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM

- neural networks. *Energy*, 182, 72–81. <https://doi.org/10.1016/j.energy.2019.05.230>.
- Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, 18(2). <https://doi.org/10.3390/s18020401>.
- Ko, B. C., Jeong, M., & Nam, J. (2014). Fast human detection for intelligent monitoring using surveillance visible sensors. *Sensors (Switzerland)*, 14(11), 21247–21257. <https://doi.org/10.3390/s141121247>.
- Ko, B. C., Kim, D.-Y., Jung, J.-H., & Nam, J.-Y. (2013). Three-level cascade of random forests for rapid human detection. *Optical Engineering*, 52(2), 027204. <https://doi.org/10.1117/1.oe.52.2.027204>.
- Kotsia, I., Buciu, I., & Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7), 1052–1067. <https://doi.org/10.1016/j.imavis.2007.11.004>.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25. <http://code.google.com/p/cuda-convnet/>.
- Kuramoto, E., Yoshinaga, S., Nakao, H., Nemoto, S., & Ishida, Y. (2019). Characteristics of facial muscle activity during voluntary facial expressions: Imaging analysis of facial expressions based on myogenic potential data. *Neuropsychopharmacology Reports*, 39(3), 183–193. <https://doi.org/10.1002/npr2.12059>.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite BEDRT for self-supervised learning of language representations. In *The International Conference on Learning Representations (ICLR)*.
- Latreche, A., Kelaiaia, R., Chemori, A., & Kerboua, A. (2023). Reliability and validity analysis of MediaPipe-based measurement system for some human rehabilitation motions. *Measurement: Journal of the International Measurement Confederation*, 214. <https://doi.org/10.1016/j.measurement.2023.112826>.
- Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., & Bottou, E. (1998). Gradient-based Learning Applied To Document Recognition - Proceedings of the IEEE. *Proceedings of the IEEE*, 86(11). <https://doi.org/10.1109/5.726791i>.
- LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual review of neuroscience*, 23(1), 155-184.
- Lee, S., Yoon, H., Park, S., Lee, S., & Kang, J. (2023). Stabilized Temporal 3D Face Alignment Using Landmark Displacement Learning. *Electronics*, 12(17), 3735.
- Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., & Huang, F. (2019). DSFD: Dual Shot Face Detector. *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, 5060–5069.

- Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikäinen, M. (2013). A Spontaneous Micro-expression Database: Inducement, Collection and Baseline. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (Fg), 1–6.
- Li, Y., Wei, J., Liu, Y., Kauttonen, J., & Zhao, G. (2022). Deep Learning for Micro-Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 13(4), 2028–2046. <https://doi.org/10.1109/TAFFC.2022.3205170>.
- Liang, Y., Hao, Y., Liao, J., Deng, Z., Wen, X., Zheng, Z., & Pan, J. (2023). A spatiotemporal network using a local spatial difference stack block for facial micro-expression recognition. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-023-16033-1>.
- Lin, Z., Li, M., Zheng, Z., Cheng, Y., & Yuan, C. (2020). Self-Attention ConvLSTM for Spatiotemporal Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11531–11538. www.aaai.org
- Liu, M., Li, S., Shan, S., & Chen, X. (2015). AU-inspired Deep Networks for Facial Expression Feature Learning. *Neurocomputing*, 159(1), 126–136. <https://doi.org/10.1016/j.neucom.2015.02.011>.
- Liu, M., Li, S., Shan, S., Wang, R., & Chen, X. (2015). Deeply learning deformable facial action parts model for dynamic expression analysis. *Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12*. Springer International Publishing., 143–157. <https://doi.org/10.1007/978-3-319-16817-3>.
- Liu, R., Wang, X., Liu, J., & Zhou, J. (2024, August). A comprehensive analysis of evaluating robustness and generalization ability of models in AES. In *Journal of Physics: Conference Series* (Vol. 2813, No. 1, p. 012022). IOP Publishing.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Lu, W., Rui, H., Liang, C., Jiang, L., Zhao, S., & Li, K. (2020). A method based on GA-CNN-LSTM for daily tourist flow prediction at scenic spots. *Entropy*, 22(3). <https://doi.org/10.3390/e22030261>
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. 2010 Ieee Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. *IEEE.*, 94–101.

- Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, 57–64. <https://doi.org/10.1109/FG.2011.5771462>
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines. <http://arxiv.org/abs/1906.08172>
- Luo, Y., Wu, C. M., & Zhang, Y. (2013). Facial expression recognition based on fusion feature of PCA and LBP with SVM. *Optik*, 124(17), 2767–2770. <https://doi.org/10.1016/j.ijleo.2012.08.040>
- Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding Facial Expressions with Gabor Wavelets. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE.*, 200–205.
- Madrigal-Garcia, M. I., Rodrigues, M., Shenfield, A., Singer, M., Cuesta, J., Jeronimo, D., Cuesta, M., & Moreno Cuesta, J. (2018). Title What faces reveal: a novel method to identify patients at risk of deterioration using facial expressions. *Critical Care Medicine*, 46(7), 1057–1062.
- Malik, Y. S., Sabahat, N., & Moazzam, M. O. (2020, November 5). Image Animations on Driving Videos with DeepFakes and Detecting DeepFakes Generated Animations. *Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020*. <https://doi.org/10.1109/INMIC50486.2020.9318064>
- Manalu, H. V., & Rifai, A. P. (2024). Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intelligent Systems with Applications*, 21. <https://doi.org/10.1016/j.iswa.2024.200339>
- Meenal, R., & Selvakumar, A. I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. *Renewable Energy*, 121, 324–343. <https://doi.org/10.1016/j.renene.2017.12.005>
- Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences*, 2(3). <https://doi.org/10.1007/s42452-020-2234-1>
- Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., & Wang, L. (2019). A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. *Journal of Hydrology*, 568, 462–478. <https://doi.org/10.1016/j.jhydrol.2018.11.015>
- Miyoshi, R., Nagata, N., & Hashimoto, M. (2021). Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video. *Neural Computing and Applications*, 33(13), 7381–7392. <https://doi.org/10.1007/s00521-020->

- Mohan, K., Seal, A., Krejcar, O., & Yazidi, A. (2021). Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks. *IEEE Transactions on Instrumentation and Measurement*, 70. <https://doi.org/10.1109/TIM.2020.3031835>.
- Mohan, P., Ravichandran, T., Prakash, M., & Ravichandran, T. (2012). An efficient resource selection and binding model for job scheduling in grid. In *European Journal of Scientific Research* (Vol. 81, Issue 4). <http://www.europeanjournalofscientificresearch.com>.
- Moishin, M., Deo, R. C., Prasad, R., Raj, N., & Abdulla, S. (2021). Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm. *IEEE Access*, 9, 50982–50993. <https://doi.org/10.1109/ACCESS.2021.3065939>.
- Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going Deeper in Facial Expression Recognition using Deep Neural Networks. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE., 1–10.
- Mourao, A., & Magalhaes, J. (2013). Competitive affective gaming: Winning with a smile. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 83–92. <https://doi.org/10.1145/2502081.2502115>.
- Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, 5(3).
- Mukashev, D., Kairgaliyev, M., Alibekov, U., Oralbayeva, N., & Sandygulova, A. (2021). Facial expression generation of 3D avatar based on semantic analysis. 2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021, 89–94. <https://doi.org/10.1109/RO-MAN50785.2021.9515463>.
- Mukhiddinov, M., Djuraev, O., Akhmedov, F., Mukhamadiyev, A., & Cho, J. (2023). Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People. *Sensors*, 23(3). <https://doi.org/10.3390/s23031080>.
- Nagireddi, J. N., Vyas, A. K., Sanapati, M. R., Soin, A., & Manchikanti, L. (2022). The Analysis of Pain Research through the Lens of Artificial Intelligence and Machine Learning. *Pain Physician*. American Society of Interventional Pain Physician, 25(2), E211–E243. www.painphysicianjournal.com.
- Nazarkevych, M., Lutsyshyn, V., Nazarkevych, H., Parkhuts, L., & Kostiak, M. (2023). Methods of Face Recognition in Video Sequences and Performance Studies. *CEUR Workshop Proceedings* (Vol. 3421, pp. 246-253). <https://CEUR-WS.org>.
- Niu, X., Han, H., Yang, S., Huang, Y., & Shan, S. (2019). Local Relationship Learning with Person-specific Shape Regularization for Facial Action Unit Detection. *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 11917–11926.

- Niu, X. X., & Suen, C. Y. (2012). A novel hybrid CNN-SVM classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4), 1318–1325. <https://doi.org/10.1016/j.patcog.2011.09.021>.
- Noyes, E., Davis, J. P., Petrov, N., Gray, K. L. H., & Ritchie, K. L. (2021). The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science*, 8(3). <https://doi.org/10.1098/rsos.201169>.
- Odell, M., Victor, C., & Oliver, D. (2009). Nurses' role in detecting deterioration in ward patients: Systematic literature review. In *Journal of Advanced Nursing* (Vol. 65, Issue 10, pp. 1992–2006). <https://doi.org/10.1111/j.1365-2648.2009.05109.x>.
- Orrite, C., Gañán, A., & Rogez, G. (2009). HOG-based decision tree for facial expression classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5524 LNCS, 176–183. https://doi.org/10.1007/978-3-642-02172-5_24.
- Ouyang, Y., & Sang, N. (2013). Robust automatic facial expression detection method. *Journal of Software*, 8(7), 1759–1764. <https://doi.org/10.4304/jsw.8.7.1759-1764>.
- Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005a). WEB-BASED DATABASE FOR FACIAL EXPRESSION ANALYSIS. 2005 IEEE International Conference on Multimedia and Expo. IEEE., 5.
- Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005b). WEB-BASED DATABASE FOR FACIAL EXPRESSION ANALYSIS. IEEE International Conference on Multimedia and Expo. IEEE.
- Pazhoohi, F., Forby, L., & Kingstone, A. (2021). Facial masks affect emotion recognition in the general population and individuals with autistic traits. *PLoS ONE*, 16(9 September). <https://doi.org/10.1371/journal.pone.0257740>.
- Phan, H., Hertel, L., Maass, M., & Mertins, A. (2016). Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks. <http://arxiv.org/abs/1604.06338>.
- Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, 16(2), 331-348.
- Peralta, D., & Saeys, Y. (2020). Robust unsupervised dimensionality reduction based on feature clustering for single-cell imaging data. *Applied Soft Computing*, 93, 106421.
- Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature medicine*, 25(1), 37-43.
- Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2), 267–274.

<https://doi.org/10.1016/j.pain.2008.04.010>.

- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Preprint. 1–12.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.
- Rani, A., Kumar, N., Kumar, J., & Sinha, N. K. (2022). Machine learning for soil moisture assessment. In *Deep Learning for Sustainable Agriculture* (pp. 143–168). Elsevier. <https://doi.org/10.1016/B978-0-323-85214-2.00001-X>.
- Raouhi, E. M., Lachgar, M., Hrimech, H., & Kartit, A. (2022). Optimization techniques in deep convolutional neuronal networks applied to olive diseases classification. *Artificial Intelligence in Agriculture*, 6, 77–89. <https://doi.org/10.1016/j.aiia.2022.06.001>.
- Ray, S. (2019). A quick review of machine learning algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE., 35–39.
- Reza, M. A. N., Hamidi, E. A. Z., Ismail, N., Effendi, M. R., Mulyana, E., & Shalannanda, W. (2021). Design a Landmark Facial-Based Drowsiness Detection Using Dlib And Opencv For Four-Wheeled Vehicle Drivers. 2021 15th International Conference on Telecommunication Systems, Services, and Applications (TSSA), 1–5.
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1), 1-7.
- 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. <https://doi.org/10.1037/0033-2909.95.1.52>.
- Roberson, D., Kikutani, M., Döge, P., Whitaker, L., & Majid, A. (2012). Shades of emotion: What the addition of sunglasses or masks to faces reveals about the development of facial expression processing. *Cognition*, 125(2), 195–206. <https://doi.org/10.1016/j.cognition.2012.06.018>.
- Rudovic, O., Pavlovic, V., & Pantic, M. (2015). Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5), 944–958. <https://doi.org/10.1109/TPAMI.2014.2356192>.
- Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning representations by back-propagation errors. *Nature*, 323(6088), 533–536.
- Safavi, F., Patel, K., & Vinjamuri, R. K. (2023). Towards Efficient Deep Learning Models for Facial Expression Recognition using Transformers. 2023 IEEE 19th International Conference on Body Sensor Networks, BSN 2023 - Proceedings.

<https://doi.org/10.1109/BSN58485.2023.10331041>.

- Sang, D. V., Dat, N. Van, & Do, P. T. (2017). Facial Expression Recognition Using Deep Convolutional Neural Networks. 9th International Conference on Knowledge and Systems Engineering (KSE), 130–135.
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3), 160.
- Sato, W., Hyniewska, S., Minemoto, K., & Yoshikawa, S. (2019). Facial expressions of basic emotions in Japanese laypeople. *Frontiers in Psychology*, 10(FEB). <https://doi.org/10.3389/fpsyg.2019.00259>.
- Savin, A. V., Sablina, V. A. ;, & Nikiforov, M. B. (2021). Comparison of facial landmark detection methods for micro-expressions analysis. 2021 10th Mediterranean Conference on Embedded Computing (MECO). IEEE., 1–4.
- Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? In *Emotion* (Vol. 7, Issue 1, pp. 113–130). <https://doi.org/10.1037/1528-3542.7.1.113>.
- Schmarje, L., Santarossa, M., Schröder, S. M., & Koch, R. (2021). A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 9, 82146-82168.
- Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Franconeri, S. L. (2014). Eye movements during emotion recognition in faces. *Journal of Vision*, 14(13), 1–16. <https://doi.org/10.1167/14.13>.
- Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, 27(6), 803–816. <https://doi.org/10.1016/j.imavis.2008.08.005>.
- Sharma, P. (2022). Spontaneous Facial Micro Expression Recognition and Analysis using Varying Resolutions. Ulster University.
- Sharma, U., Faisal, K. N., Sharma, R. R., & Arya, K. V. (2023). Facial Landmark-Based Human Emotion Recognition Technique for Oriented Viewpoints in the Presence of Facial Attributes. *SN Computer Science*, 4(3). <https://doi.org/10.1007/s42979-023-01727-y>.
- Sheremet, O., I., Sadovoi, O., V., Harshanov, D., V., Kovalchuk, O., S., Sheremet, K., S., & Sokhina, T., V. (2023). Efficient face detection and replacement in the creation of simple fake videos. *Applied Aspects of Information Technology*, 6(3), 286–303. <https://doi.org/10.15276/aait.06.2023.20>.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Kong Observatory, H.

- (2015a). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28.
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Kong Observatory, H. (2015b). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28.
- Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2020). First Order Motion Model for Image Animation. <http://arxiv.org/abs/2003.00196>.
- Singh, A. K., Kumbhare, V. A., & Arthi, K. (2021). Real-Time Human Pose Detection and Recognition Using MediaPipe. *International Conference on Soft Computing and Signal Processing*, 145–154.
- Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S. (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology (Singapore)*, 15(4), 1819–1830. <https://doi.org/10.1007/s41870-023-01183-0>.
- Sitepu, S. E., Jati, G., Alhamidi, M. R., Caesarendra, W., & Jatmiko, W. (2021). FaceNet with RetinaFace to Identify Masked Face. *Proceedings - IWBIS 2021: 6th International Workshop on Big Data and Information Security*, 81–86. <https://doi.org/10.1109/IWBIS53353.2021.9631848>.
- Smith, M. L., Cottrell, G. W., Dé Ric Gosselin, F., & Schyns, P. G. (2005). Transmitting and Decoding Facial Expressions (Vol. 16, Issue 3). <http://www.cs.ucsd.edu/users/gary/>.
- Sollfrank, T., Kohlen, O., Hilfiker, P., Kegel, L. C., Jokeit, H., Brugger, P., Loertscher, M. L., Rey, A., Mersch, D., Sternagel, J., Weber, M., & Grunwald, T. (2021). The Effects of Dynamic and Static Emotional Facial Expressions of Humans and Their Avatars on the EEG: An ERP and ERD/ERS Study. *Frontiers in Neuroscience*, 15. <https://doi.org/10.3389/fnins.2021.651044>.
- Sorokin, M., Zhdanov, D., & Zhdanov, A. (2018). Recovery of optical parameters of a scene using fully-convolutional neural networks. *MICSECS*.
- Street, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Education and Counseling*, 74(3), 295–301. <https://doi.org/10.1016/j.pec.2008.11.015>.
- Suk, M., & Prabhakaran, B. (2014). Real-time Mobile Facial Expression Recognition System-A Case Study. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 132–137.
- Sukawattanavijit, C., Chen, J., & Zhang, H. (2017). GA-SVM Algorithm for Improving Land-Cover Classification Using SAR and Optical Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 14(3), 284–288. <https://doi.org/10.1109/LGRS.2016.2628406>.

- Sun, B., Cao, S., Li, D., He, J., & Yu, L. (2022). Dynamic Micro-Expression Recognition Using Knowledge Distillation. *IEEE Transactions on Affective Computing*, 13(2), 1037–1043. <https://doi.org/10.1109/TAFFC.2020.2986962>.
- Sun, Y., Xue, B., Zhang, M., Yen, G. G., & Lv, J. (2020). Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Transactions on Cybernetics*, 50(9), 3840–3854. <https://doi.org/10.1109/TCYB.2020.2983860>.
- Sutanto, D. I., Verine, Salim, M. V., & Ham, H. (2021). Auto-Tracking Camera System for Remote Learning Using Face Detection and Hand Gesture Recognition Based on Convolutional Neural Network. *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*, 451–457. <https://doi.org/10.1109/ICCSAI53272.2021.9609744>.
- Tanchotsrinon, C., Phimoltares, S., & Maneeroj, S. (2011). Facial expression recognition using graph-based features and artificial neural networks. *IEEE. 2011 IEEE International Conference on Imaging Systems and Techniques*, 331–334.
- Tang, X., Du, D. K., He, Z., & Liu, J. (2018). PyramidBox: A Context-assisted Single Shot Face Detector. *Proceedings of the European Conference on Computer Vision (ECCV)*, 797–813. <https://github.com/PaddlePaddle/models/tree/develop/>.
- Tang, Y. (2013). Deep Learning using Linear Support Vector Machines. <http://arxiv.org/abs/1306.0239>.
- Tao, P., Sun, Z., & Sun, Z. (2018). An Improved Intrusion Detection Algorithm Based on GA and SVM. *IEEE Access*, 6, 13624–13631. <https://doi.org/10.1109/ACCESS.2018.2810198>.
- Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied Soft Computing Journal*, 75, 323–332. <https://doi.org/10.1016/j.asoc.2018.11.001>.
- Tharwat, A. (2021). Classification assessment methods. *Applied computing and informatics*, 17(1), 168-192.
- Thaseen, I. S., & Kumar, C. A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University-Computer and Information Sciences*, 29(4), 462-472.
- Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B. (2018). LSTM-based traffic flow prediction with missing data. *Neurocomputing*, 318, 297–305. <https://doi.org/10.1016/j.neucom.2018.08.067>.
- Tirupal, T., Kumar, M. N., Basha, P. M., Babu, J. M., & Rathan, O. (2023). OPENCV Based Smart Attendance System Using Facial Recognition. *2023 4th International Conference for Emerging Technology, INCET 2023*. <https://doi.org/10.1109/INCET57972.2023.10170456>.

- Treal, T., Jackson, P. L., & Meugnot, A. (2020). Combining trunk movement and facial expression enhances the perceived intensity and believability of an avatar's pain expression. *Computers in Human Behavior*, 112. <https://doi.org/10.1016/j.chb.2020.106451>.
- Tsai, C.-F., Lin, W.-Y., Hong, Z.-F., & Hsieh, C.-Y. (2011). Distance-based features in pattern classification. *EURASIP Journal on Advances in Signal Processing*, 2011(1). <https://doi.org/10.1186/1687-6180-2011-62>.
- Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.
- Uddin, M. Z., Lee, J. J., & Kim, T.-S. (2009). An Enhanced Independent Component-Based Human Facial Expression Recognition from Video. In *IEEE Transactions on Consumer Electronics* (Vol. 55, Issue 4).
- Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise An addition to the mmi facial expression. *Proc. 3rd Intern. Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect*. Paris, France., 10, 65–70.
- Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 9. Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.
- Vats, A., & Chadha, A. (2023). Facial Expression Recognition using Squeeze and Excitation-powered Swin Transformers. <http://arxiv.org/abs/2301.10906>.
- Viola, P., & Jones, M. (2001). Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade. *Advances in Neural Information Processing Systems*, 14.
- Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*. CVPR 2001 (Vol. 1, pp. I-I). IEEE.
- Wang, F., Zhen, Z., Wang, B., & Mi, Z. (2017). Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting. *Applied Sciences (Switzerland)*, 8(1), 28. <https://doi.org/10.3390/app8010028>.
- Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing*, 6(1), 69–75. <https://doi.org/10.1109/TAFFC.2015.2392101>.
- Wang, Y., Ai, H., Wu, B., & Huang, C. (2004). Real Time Facial Expression Recognition with Adaboost. *Proceedings of the 17th International Conference on Pattern Recognition*, 2004.

ICPR 2004, 3, 926–929.

- Wang, Y., Gan, D., Sun, M., Zhang, N., Lu, Z., & Kang, C. (2019). Probabilistic individual load forecasting using pinball loss guided LSTM. *Applied Energy*, 235, 10–20. <https://doi.org/10.1016/j.apenergy.2018.10.078>.
- Wang, Y., Gao, Z., Long, M., Wang, J., & Yu, P. S. (2018). PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning.
- Wang, Y., Li, Y., Song, Y., & Rong, X. (2020). The influence of the activation function in a convolution neural network model of facial expression recognition. *Applied Sciences (Switzerland)*, 10(5). <https://doi.org/10.3390/app10051897>.
- Wang, Y., Long, M., Wang, J., Gao, Z., & Yu, P. S. (2017). PredRNN: Recurrent Neural Networks for Predictive Learning using Spatiotemporal LSTMs. *Advances in Neural Information Processing Systems*, 30.
- Wang, Y., Zhang, J., Zhu, H., Long, M., Wang, J., Kliss, P. S. Y., & Bnrist, M. ; (2019). Memory In Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity from Spatiotemporal Dynamics Beijing Key Laboratory for Industrial Big Data System and Application. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9154–9162.
- Wood, E., Baltrusaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljevic, N., ... & Stojiljkovic, I. 3D face reconstruction with dense landmarks. *arXiv 2022*. arXiv preprint arXiv:2204.02776, 3.
- Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, 171, 109203.
- Wu, Y., & Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2), 115-142.
- Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PLoS ONE*, 12(5). <https://doi.org/10.1371/journal.pone.0177239>.
- Xia, B., Wang, W., Wang, S., & Chen, E. (2020, October). Learning from macro-expression: A micro-expression recognition framework. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2936-2944).
- Xie, H. X., Lo, L., Shuai, H. H., & Cheng, W. H. (2020). AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 2871–2880. <https://doi.org/10.1145/3394171.3414012>.

- Xu, X., & De Sa, V. R. (2020). Exploring Multidimensional Measurements for Pain Evaluation using Facial Action Units. *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, 786–792. <https://doi.org/10.1109/FG47880.2020.00087>.
- Xue, D. X., Zhang, R., Feng, H., & Wang, Y. L. (2016). CNN-SVM for Microvascular Morphological Type Recognition with Data Augmentation. *Journal of Medical and Biological Engineering*, 36(6), 755–764. <https://doi.org/10.1007/s40846-016-0182-4>.
- Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., Zong, Y., & Sun, N. (2016). Multi-clue fusion for emotion recognition in the wild. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 458–463. <https://doi.org/10.1145/2993148.2997630>.
- Yan, W. J., Wang, S. J., Liu, Y. J., Wu, Q., & Fu, X. (2014). For micro-expression recognition: Database and suggestions. *Neurocomputing*, 136, 82-87.
- Yan, W. J., Wu, Q., Liu, Y. J., Wang, S. J., & Fu, X. (2013, April). CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1-7). IEEE.
- Yan, Y., Wang, Y., Gao, W. C., Zhang, B. W., Yang, C., & Yin, X. C. (2018). LSTM 2 : Multi-Label Ranking for Document Classification. *Neural Processing Letters*, 47(1), 117–138. <https://doi.org/10.1007/s11063-017-9636-0>.
- Yang, H., Ciftci, U., & Yin, L. (2018). Facial Expression Recognition by De-expression Residue Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2168–2177.
- Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). From Facial Parts Responses to Face Detection: A Deep Learning Approach. *Proceedings of the IEEE International Conference on Computer Vision*, 3676–3684.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q.V. (2019). XLnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*.
- Ye, C., Wang, O., Liu, M., Zheng, L., Xia, M., Hao, S., Jin, B., Jin, H., Zhu, C., JungHuang, C., Gao, P., Ellrodt, G., Brennan, D., Stearns, F., Sylvester, K. G., Widen, E., McElhinney, D. B., & Ling, X. (2019). A real-time early warning system for monitoring inpatient mortality risk: Prospective study using electronic medical record data. *Journal of Medical Internet Research*, 21(7). <https://doi.org/10.2196/13719>.
- Yuan, X., Zhang, S., Zhao, C., He, X., Ouyang, B., & Yang, S. (2022). Pain Intensity Recognition

- from Masked Facial Expressions using Swin-Transformer. 2022 IEEE International Conference on Robotics and Biomimetics, ROBIO 2022, 723–728. <https://doi.org/10.1109/ROBIO55434.2022.10011731>.
- Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138, 1-24.
- Zavaschi, T. H. H., Britto, A. S., Oliveira, L. E. S., & Koerich, A. L. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, 40(2), 646–655. <https://doi.org/10.1016/j.eswa.2012.07.074>.
- Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. <http://arxiv.org/abs/1311.2901>.
- Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2007, November). A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces* (pp. 126-133).
- Zhan, C., Li, W., Ogunbona, P., & Safaei, F. (2008). A Real-Time Facial Expression Recognition System for Online Games. *International Journal of Computer Games Technology*, 2008, 1–7. <https://doi.org/10.1155/2008/542918>.
- Zhang, F., Zhang, T., Mao, Q., & Xu, C. (2018). Joint Pose and Expression Modeling for Facial Expression Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3359–3368.
- Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503. <https://doi.org/10.1109/LSP.2016.2603342>.
- Zhang, L., Zhu, G., Mei, L., Shen, P., Afaq, S., Shah, A., & Bennamoun, M. (2018). Attention in Convolutional LSTM for Gesture Recognition. *Advances in Neural Information Processing Systems*, 31. <https://github.com/GuangmingZhu/AttentionConvLSTM>.
- Zhang, S., Zhang, S., Wang, B., & Habetler, T. G. (2020). Deep learning algorithms for bearing fault diagnostics—A comprehensive review. *IEEE Access*, 8, 29857-29881.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). S³FD: Single Shot Scale-invariant Face Detector. *Proceedings of the IEEE International Conference on Computer Vision*, 192–201.
- Zhang, X., Yu, Y., Wang, L., & Gu, Q. (2019, April). Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics* (pp. 1524-1534). PMLR.
- Zhang, Y. (2024). A Review of Deep Learning-Based Micro-Expression Classification. *Highlights*


in Science, Engineering and Technology, 103, 109-114.

- Zhang, Y., Zhao, Y., Wen, Y., Tang, Z., Xu, X., & Liu, M. (2021). Facial Prior Based First Order Motion Model for Micro-expression Generation. MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia, 4755–4759. <https://doi.org/10.1145/3474085.3479211>.
- Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep Region and Multi-label Learning for Facial Action Unit Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3391–3399.
- Zhao, R., Yan, R., Wang, J., & Mao, K. (2017). Learning to monitor machine health with convolutional Bi-directional LSTM networks. Sensors (Switzerland), 17(2). <https://doi.org/10.3390/s17020273>.
- Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. ACM computing surveys (CSUR), 35(4), 399-458.
- Zhao, X., Liu, Y., Chen, T., Wang, S., Chen, J., Wang, L., & Liu, G. (2022). Differences in brain activations between micro- and macro-expressions based on electroencephalography. Frontiers in Neuroscience, 16. <https://doi.org/10.3389/fnins.2022.903448>.
- Zheng, Y., & Blasch, E. (2023). Facial Micro-Expression Recognition Enhanced by Score Fusion and a Hybrid Model from Convolutional LSTM and Vision Transformer. Sensors, 23(12). <https://doi.org/10.3390/s23125650>.
- Zhi, R., Liu, M., Xu, H., & Wan, M. (2019). Facial micro-expression recognition using enhanced temporal feature-wise model. Communications in Computer and Information Science, 1138 CCIS, 301–311. https://doi.org/10.1007/978-981-15-1925-3_22.
- Zhou, C., Sun, C., Liu, Z., & Lau, F. C. M. (2015). A C-LSTM Neural Network for Text Classification. ArXiv Preprint ArXiv:1511.08630. <http://arxiv.org/abs/1511.08630>
- Zhou, X., Wan, X., & Xiao, J. (2016). Attention-based LSTM Network for Cross-Lingual Sentiment Classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 247–256.
- Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2020). Gradient descent optimizes over-parameterized deep ReLU networks. Machine Learning, 109(3), 467–492. <https://doi.org/10.1007/s10994-019-05839-6>.

APPENDIX A. PUBLISHED PAPER

Article

AI-Based Visual Early Warning System

Zeena Al-Tekreeti ^{1,*}, Jeronimo Moreno-Cuesta ², Maria Isabel Madrigal Garcia ² and Marcos A. Rodrigues ¹ 

¹ Industry and Innovation Research Institute, Sheffield Hallam University, Sheffield S1 1WB, UK; m.rodrigues@shu.ac.uk

² Department of Intensive Care, North Middlesex University Hospital, London N18 1QX, UK; jeronimo.moreno-cuesta@nhs.net (J.M.-C.); maria.madrigal@nhs.net (M.I.M.G.)

* Correspondence: b3036983@hallam.shu.ac.uk

Abstract: Facial expressions are a universally recognised means of conveying internal emotional states across diverse human cultural and ethnic groups. Recent advances in understanding people's emotions expressed through verbal and non-verbal communication are particularly noteworthy in the clinical context for the assessment of patients' health and well-being. Facial expression recognition (FER) plays an important and vital role in health care, providing communication with a patient's feelings and allowing the assessment and monitoring of mental and physical health conditions. This paper shows that automatic machine learning methods can predict health deterioration accurately and robustly, independent of human subjective assessment. The prior work of this paper is to discover the early signs of deteriorating health that align with the principles of preventive reactions, improving health outcomes and human survival, and promoting overall health and well-being. Therefore, methods are developed to create a facial database mimicking the underlying muscular structure of the face, whose Action Unit motions can then be transferred to human face images, thus displaying animated expressions of interest. Then, building and developing an automatic system based on convolution neural networks (CNN) and long short-term memory (LSTM) to recognise patterns of facial expressions with a focus on patients at risk of deterioration in hospital wards. This research presents state-of-the-art results on generating and modelling synthetic database and automated deterioration prediction through FEs with 99.89% accuracy. The main contributions to knowledge from this paper can be summarized as (1) the generation of visual datasets mimicking real-life samples of facial expressions indicating health deterioration, (2) improvement of the understanding and communication with patients at risk of deterioration through facial expression analysis, and (3) development of a state-of-the-art model to recognize such facial expressions using a ConvLSTM model.



Citation: Al-Tekreeti, Z.;

Moreno-Cuesta, J.; Madrigal Garcia,

M.I.; Rodrigues, M.A. AI-Based Visual

Early Warning System. *Informatics*

2024, 11, 59. [https://doi.org/](https://doi.org/10.3390/informatics11030059)

10.3390/informatics11030059

Received: 8 April 2024

Revised: 1 July 2024

Accepted: 16 July 2024

Published: 12 August 2024



Copyright: © 2024 by the authors.

Licensed under MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the Creative Commons

Attribution (CC BY) license ([https://](https://creativecommons.org/licenses/by/4.0/)

[creativecommons.org/licenses/by/](https://creativecommons.org/licenses/by/4.0/)

4.0/).

Keywords: facial expression (FE); facial expression recognition (FER); automatic facial expression recognition (AFER); machine learning (ML); deep learning (DL); convolution neural networks (CNN); long short-term memory (LSTM)

1. Introduction

An understanding of human feelings, behaviours, and intentions is based on interpreting their sentiments expressed by various cues, including verbal communication, such as speech patterns, and nonverbal communication, such as body language, gestures, head nods, and facial expressions (FEs) [1]. The prefrontal cortex, limbic system, and hippocampus are responsible for processing human feelings and emotional messages [2,3]. In the healthcare field, communication is an essential factor for understanding patient health and well-being [4]. However, it can be difficult to communicate directly with patients in critical states for a number of reasons, such as unconsciousness, severe illness, inability to speak, under medication, cognitive impairment, mental disorder, or second language. In such cases, a medical team must rely on measurements like vital signs monitoring (including heart rate (HR), blood pressure (BP), temperature, oxygen saturation), pain assessment

APPENDIX A. PUBLISHED PAPER

(using the visual analogue scale), and imaging tests (including chest X-rays, CT-scans, ultrasound). Normally, facial expressions are not considered in these situations, and this is the aim of this research. Deterioration is a critical state and a serious factor that may result in the death of the patient [5] and can impair patients' ability to communicate and convey their feelings, thoughts, and desires [6]. The early warning clinical signs of severe deterioration are important indicators that emphasize the imperative of taking preventive measures and immediate proper actions to improve patient health and increase their chance of survival [7]. An early and accurate clinical assessment, which is usually checked by professional nurses, is known as a self-report and is not always possible due to some factors such as age, the critical situation of the patient, unconsciousness, language impairments, the ability to speak, and difficulty in explaining their sentiments [8]. Self-report is a costly procedure, time-consuming concerning human resources, and difficult to perform objectively. In addition, it can be a risk assessment by the critical care nurses in intensive care units who depend on direct evaluation because of the likelihood of being alerted by their intuitive decisions [3,9]. Therefore, there is a need for an automatic system that can perform accurate measurements and health assessments, providing an early warning and prioritizing patients in need of urgent medical care. Face is the organ of emotion which is considered an essential indicator of human feelings, the prime source of wealth information, and the significant channel for transferring nonverbal communication [10]. Psychologists have illustrated facial expressions form 55% of daily human interaction, which is considered far higher than other verbal communication, such as speech-language, with 38%, and the written language, with only 7% [11]. Facial expressions are formed by the coordination of facial muscles that play a primary role in the exchange of information and facilitating social interactions [12]. The facial muscle movements are stimulated by the facial nerves, forming various types of voluntary and involuntary facial expressions. Facial analysis is a vital domain to multiple aspects of daily life such as age estimation, gender classification, face detection, face recognition, face posing, facial expression, and blink detection. Consequently, facial image analysis has been used in healthcare disciplines such as pain estimation, psychological assessment, and the analysis of mental health. All the achieved exceptional outcomes and continued progress in the facial analysis field have led to demonstrate its usefulness and effectiveness, which are reflected in various disciplines and applications. Some movements of facial muscles that surround all main facial landmarks, such as the eyes, nose, mouth, and ears, form specific expressions called universal facial expressions, which are perceived as happiness, anger, contempt, disgust, surprise, fear, and sadness. Other emotions reveal human sentiments and state of mind, allowing other people to glimpse into human minds as they can read facial expressions based on changes in key facial features. In 1978, Ekman and Friesen [13] developed a system to characterize facial expressions, which is called The Facial Action Coding System (FACS). This system refers to a comprehensive set of atomic non-overlapping facial muscle actions called Action Units (AUs) that are typically used to encode and taxonomize facial muscle movements to correspond to a displayed emotion [14]. Recently, different methods of facial expression recognition based on FACS have been used to identify the seven universal facial emotions such as joy, sadness, contempt, surprise, disgust, fear, and anger. However, human facial expression consists of thousands of expressions that are different in subtle changes due to a variety of Facial Action Units (FAUs), or the blending of some expressions. FACS can identify large numbers of facial emotions by identifying a set of muscular movements that comprise the facial expression. In facial behaviour, there are different relationships among FAUs. For instance, a set of AUs usually appears together to show specific emotions, such as the co-occurrence relationship of inner brow raiser (AU1) and outer brow raiser (AU2), and mutually exclusive relationships of lip presser (AU24) and lips apart (AU25). In 2015, Rudovic et al. [14] stated that identifying the estimation of facial action unit intensity is a challenging task due to some internal and external factors and conditions such as head position, illumination, age, or a specific set of action units. In the healthcare field, health professionals observe and assess facial expressions as a way to

APPENDIX A. PUBLISHED PAPER

deepen their understanding of patients without causing them the exertion of saying what they feel. Therefore, some potential applications have been utilised to recognize depression, pain, and anxiety in patients [15,16]. Evaluation of facial expressions and analysing the degree of patient deterioration have relied on assessments based on the experiences that individual nurses have acquired in the course of their careers. In addition, nurses not only observe the status of the well-being of patients based on basic facial expressions but also need to assess and interpret complex changes exhibited through the face. Furthermore, patient monitoring is based on observation by nursing staff, which means that such measurements and patient deterioration may not be noticed in the time between observations. Moreover, the effort of training human experts and manually scoring the AUs is expensive and time-consuming. Within the realm of computer vision, Facial Expression Recognition (FER) stands as a crucial field, offering diverse techniques to decode emotions from facial expressions. Navigating the human-machine interaction (HMI) has a significant impact on feeling, recognising, and understanding internal emotions and intentions. Nowadays, there is a remarkable steady growth in employing digital images and machine learning (ML) in facial recognition and human-computer interaction due to the availability of high-end devices such as image devices (cameras) and cost-effectiveness. Facial image detection, analysis, and recognition have evolved into a substantial work that ended with remarkable outcomes. The main concept of involving deep learning algorithms is to achieve certain requirements by constructing a robust artificial neural network (ANN) model through training an enormous quantity of datasets along with considering their diversity and quality to satisfy certain requirements [17]. In 2022, Rodriguez and his colleagues [18] stated that the automatic recognition of deterioration is an essential part of the health domain since it is not only an influential indicator for medical diagnosis but has also been shown to be a supportive factor for patient recuperation in intensive care units, admitted to critical units and after surgery. Hence, precise deterioration assessment could be highly beneficial from the early warning automatic system [3]. Consequently, more intensive monitoring and accurate methods for observing and understanding the changes in facial expressions can help to identify the risk of clinical deterioration earlier than statistical methods. This paper presents an Automatic Facial Expression Recognition (AFER) to support healthcare professionals by providing specific information revealing significant patients' health status without requiring previous knowledge or special skills. There are various feature extraction approaches aimed at providing features related to fine details of a dataset. Choosing one or multiple features is influenced by several factors such as what the specific targets or task requirements need to be achieved, the characteristics or nature of the dataset, and the dimensionality between the input dataset and targets. In this paper, to capture various aspects of facial data, a combination of appropriate feature extraction approaches has been employed for a comprehensive, robust analysis with high levels of detection 99.8% and recognition of a patient's health status. The method proposed here includes, for instance, various metrics between facial landmarks. Distances or ratio measurements between facial action units and facial landmarks were used as facial features. Providing such features together with the spatial relationships between facial landmarks to deep learning systems led to vastly improved classifier models. In particular, here, we investigated and proposed a Convolution LSTM model to learn and predict facial features from animated characters, created using special software such as Blender 4.1 and First Order Motion Model (FOMM). Guided by the work of Madrigal and her colleagues [3] whose work predicted health conditions by recognizing early signs of deterioration, our research focused on convolution neural networks and Long-Short Term Memory in an attempt to replicate such predictions from the detection and recognition of facial expressions from a set of Action Units. Thus, the work reported here concerns the development of an Automatic Facial Action Units Recognition system capable of measuring the risk of patient deterioration in critical care wards. The expressed emotional states of patients were detected in real-time using fully automated computer algorithms that receive the data of facial expressions via webcam.

APPENDIX A. PUBLISHED PAPER

2. Related Work

Human facial expressions serve as a fundamental mode of communication and interaction. Therefore, facial expression recognition (FER) is an essential part of human communication and plays an energetic role in expressing emotions and providing non-verbal cues [19]. Charles Darwin investigated facial expressions and stated that facial movements convey what we are feeling, even though interpretations may differ among cultural groups. Some researchers have investigated the various muscle movements of facial expressions and have proposed that humans display universal facial expressions for specific emotions [20]. In 1971, one of the earliest works on facial expression was presented by Ekman and his colleagues [21]. They developed their theory based on facial expressions by observing films of social human interactions in different cultures. They proposed that people have universal facial expressions for specific emotions based on analysing and recognizing data from different cultures. Seven years later, Ekman and Friesen [13] determined universal facial expressions by proposing the Facial Action Coding System, which can determine AUs that represent muscular movements. In 1995, Gosselin and his colleagues [22] implemented an experiment that included six participants from Canada to present emotions based on scenarios corresponding to six types of facial expressions. The outcomes of the Facial Action Coding System (FACS) of the presented emotions revealed that some of the theoretically predicted Action Units appeared frequently, such as AU 6 and AU 12 in happy expression, while other AUs were rarely observed, such as AU 9 in disgusting expression. Furthermore, several non-predicted AUs were observed frequently in most facial expressions. Later, Scherer and Ellgring (2007) [23] implemented their experiment by asking professional actors ($n = 12$) in Germany to present emotions based on scenarios corresponding to various ranges of facial expressions. According to the FACS analyses for the presented emotions, the outcomes of the experiment did not prove the existence of a large number of theoretically predicted AUs of basic and non-basic emotions. Thus, in recent decades, the aim of using computer vision as an essential assistance for professional healthcare people has been addressed; for instance, in 2011, Lucey and her colleagues [24] built a UNBC-McMaster database containing 200 video streams taken from 25 patients who were suffering from shoulder pain. The frames were labelled depending on the work of Prkachin and Solomon [25]. The metric is based on the Facial Action Coding System (FACS) that has been presented by Ekman, Friesen, and Hager (2002) [26], which codes different facial muscle movements with various intensity levels. Sometimes, the dataset has been considered challenging data in the subject of facial expression recognition even for clinical professionals to determine what the patient feels. So, the UNBC-McMaster Painful dataset has been used to propose new models for facial pain detection. Lucey and her colleagues (2011) [24] published baseline results with the dataset that used support vector machines SVM/AAM system to extract facial landmark features to predict painful action units (AUs) and the PSPI for the presence of pain. Facial AUs have been typically used to encode facial activity corresponding to different facial expressions such as pain or anger. In 2015, Rudovic and his colleagues [14] stated that the task of AU intensity estimation is very challenging, due to the high variability in facial expressions depending on the context, such as intensity of light, head poses, or various facial expression expressions. In 2013, an investigation study using FACS proposed by Gross and his colleagues [27] uncovered that health professionals usually recognize sadness and fear expressions in patients at risk of deterioration. A previous collaboration between North Middlesex University Hospital, University College of London Hospital, and the GMPR Research Group at Sheffield Hallam University proved, for the first time, that patterns of Facial Action Units can be used as predictors of admission to critical care. The study analysed some AUs related to the upper and lower face, head position, and eye position, with clinical measures collected within the National Early Warning Score (NEWS) [3]. In the last few decades, automatic facial expression recognition (FER) has been considered an essential part of various applications in human-computer interaction [28]. Therefore, it is considered a multidisciplinary research field as it is involved in many disciplines such as computer

APPENDIX A. PUBLISHED PAPER

vision, machine learning, psychology, neuroscience, and cognitive science [29]. In 2016, Jaiswal and Valstar [30] presented a combination of Convolution Neural Networks (CNN) and Bi-directional Long Short-Term Memory Networks (BLSTM) that can detect Facial Action Units. In 2017, Sang and his colleagues [31] introduced convolution neural networks that were capable of recognizing facial emotions; the output layer included seven neurons that were labelled according to seven expressions. The purpose was to classify each image as one of the universal facial expressions. One year later, Chen, Yang, Wang, and Zou (2017) [32] presented a convolution neural network that used a convolution kernel for feature extraction and a max pooling operation to minimize the dimensions of the extracted features. In this work, the proposed automatic recognition system of facial analysis was constructed to identify each facial image as one of the seven facial expressions. One of the earlier studies related to facial analysis was introduced by (Al Tasee, Jasim, 2020) [33]. They presented a CNN with the ability to perform the process of FER to label each face as one of the seven universal emotion categories that are considered in the JAFFE database. The CNN was trained with different grey-scale images, and the accuracy of the results was 100%. The work of Mohan, Seal, Krejcar, and Yazidi (2021) [34] introduced deep convolution neural networks (DCNN) for recognizing facial expressions. The proposed approach included two main parts. The first part focused on finding out local features from the human face using a gravitational force descriptor, while, in the second stage, the descriptor was fed into the DCNN model. The implementation of DCNN was applied through two stages. The first stage extracted edges, curves, and lines, while the second explored the holistic features. In summary, Facial Action Units have been employed to encode the different facial expressions corresponding to various facial motions with varying degrees of success and intrinsic model limitations. Each specific combination of AUs can form specific facial expressions such as happiness, sadness, anger, fear, and so on.

3. Methodology

Designing and developing a Convolution LSTM model involves systematic methodology to ensure effective design, training, and evaluation. This section presents three major phases that include: generating the dataset, pre-processing the dataset, and the proposed system, which is based on a Convolution LSTM architecture.

3.1. The Dataset






3.1.1. Generating the Dataset

Creating avatars using computer-generated animated characters has significantly increased over the last few years and is considered a valuable tool in studying emotions and social cognition. Using avatars provides highly controllable, interactive experiments; allows for various facial expressions; saves cost and time compared with human experiments; and encompasses a variety of data including different ages, skin tones, and ethnicities. However, there are still some limitations and drawbacks of involving avatars in research findings that have to be considered, such as that they cannot mimic or capture the richness and complexity of real facial expressions, especially the fine subtle facial movements such as micro-expressions. These shortcomings minimize the realism of avatars. Generating avatars' dynamic facial expressions emulating the risk of deterioration using Action Units was based on the Facial Action Coding System (FACS) (Ekman Friesen, 1978) [13]. Here, we used a typical combination of Action Units as described in the work of (Madrigal et al., 2018) [3] What Faces Reveal: A Novel Method to Identify Patients at Risk of Deterioration using Facial Expressions. The Action Units in question are illustrated in Table 1, where combinations of AUs are referred to Face Displays (FD). Twenty-five avatars aged between 18 and 70, with a mean age and standard deviation (Mage 30.04, SDage 14.3957), were generated with various genders, skin tones (white, black, yellow), and ethnicities (African, Asian, White), face shapes (oval, long, square, heart, diamond), and facial features (eye and hair colour, shape and size of the nose, chin, cheeks, eyes, forehead, lips). Each participant implemented five different expressions that were labelled into five classes (FD1, FD2-L,

APPENDIX A. PUBLISHED PAPER

FD2-R, FD3-L, FD3-R), representing patients whose health is deteriorating. There were 10 male participants, representing 40% of the whole avatars, while there were 15 females, representing 60% of the participants. The dataset was generated consisting of 125 video clips covering the five particular facial expressions, and each video lasted around 11–12 s, displaying the faces of avatars were created and evolved using advanced 3D animation tools such as Blender and FACSHuman. FACSHuman is a software v0.4.0 tool based on FACS, which, in conjunction with MakeHuman software 1.2.0, helps to craft 3D avatars with high standards of realism, aesthetics, and morphological precision. This involved detailed facial rigging and expressions synthesis to mimic human facial expressions accurately. The flexibility of the software allowed for the generation of FEs in limitless scenarios, enhancing the realism of synthetic data. The avatars facing the camera at various angles with a frame rate of around 25 fps showed dynamic facial expressions at risk of deterioration without the movement of the trunk. The row dataset had around 37,000 frames for the whole videos of all classes.

Table 1. Action Units of five expressions at risk of deterioration and their relevant facial muscles.

Action Unit	FACS Name	Facial Muscle	Example Image
15	Lip Corner Depressor	Depressor anguli oris (Tri-angularis)	
25	Lips part	Depressor Labii, Relax ation of Mentalis (AU17), Orbicularis Oris	
43	Eyes Closed	Relax ation of Levator Palpebrae Superioris	
55	Head Tilt Left		
56	Head Tilt Right		

The relevant Action Units of five classes and the muscles that are responsible for their appearance, along with image samples of these expressions, are shown in Table 1.

The generated avatar videos simulated actors imitating the behaviour of patients at risk of deterioration while the body was considered under static conditions and was recorded via camera, either with or without the movement of head poses. The intensity of the deterioration expression was fixed at 100% of the maximal contraction of the AUs depicted in Table 1. Each video began with the avatar showing a neutral expression (sets of AUs at 0). The level of AUs linearly increased to reach 8% of contraction for 1 s. The deterioration expression was maintained for 9 s until the end of the clip, as depicted in Figure 1. In the first few seconds (4–5 s) of each video, the avatar first stayed static along with a neutral

APPENDIX A. PUBLISHED PAPER

expression; then, the facial muscle movement started showing deterioration until reaching the maximum point at the end of the video. The automatic model analysed each frame and triggered when the participant was under deterioration risk at that captured frame. The position of the cursor along the scale was converted to numerical values between 0 ("normal situation") and 100 ("worst condition of deterioration"). The deterioration believability task rating corresponding to the percentage of "True deteriorated" responses was calculated for each condition of the five classes. The left avatar of Figure 1a expresses a neutral expression, while the right avatar reveals a deterioration condition in the final stage (AUs recruited at 90% of their maximal contraction). As illustrated in Figure 1, deterioration intensity increased significantly with the amplitude of (AU15, AU25, AU43, AU55, AU56) movement.

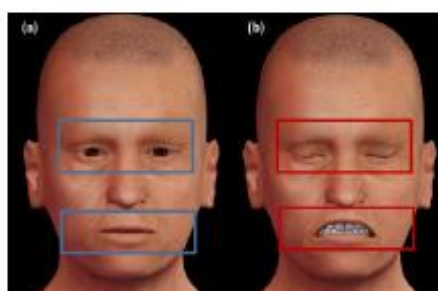


Figure 1. Facial expression areas that reveal if the patient is under deterioration or not. (a) The left avatar expresses a neutral expression, which is bounded by the blue rectangles. (b) The right avatar reveals deterioration status in the final stage, which is bounded by the red rectangles.

Table 2 shows 5 classes of different combination of AUs along with their descriptions and the number of generated videos in each class.

Table 2. Combination of AUs of each class to form facial expressions of participants at risk of deterioration and the number of generated videos.

Expressions	Involved Action Units	Description	Samples of Video Clips
FD1	AU (15 + 25 + 43)	Lip Corner Depressor, Lips part, Eyes Closed	25
FD2-L	AU (15 + 43 + 55)	Lip Corner Depressor, Eyes Closed, Head Tilt Left	25
FD2-R	AU (15 + 43 + 56)	Lip Corner Depressor, Eyes Closed, Head Tilt Right	25
FD3-L	AU (15 + 25 + 43 + 55)	Lip Corner Depressor, Lips part, Eyes Closed, Head Tilt Left	25
FD3-R	AU (15 + 25 + 43 + 56)	Lip Corner Depressor, Lips part, Eyes Closed, Head Tilt Right	25

The results showed that the avatar's deteriorated expression was perceived to be more intense and more believable in the presence of a combination of the upper part and lower part of the face. Figure 2 shows an avatar with five different expressions in perceptive of deterioration, and each particular expression belongs to one class, such as the combination of AUs (15 + 25 + 43) belonging to the class named FD1, AUs (15 + 25 + 55) belonging to the class named FD2-L, AUs (15 + 25 + 55) belonging to the class FD2-R, AUs (15 + 25 + 43 + 55) belonging to the class labelled FD3-L, and finally, the FD3-R class includes samples that show the combination of AUs (15 + 25 + 43 + 65).

APPENDIX A. PUBLISHED PAPER

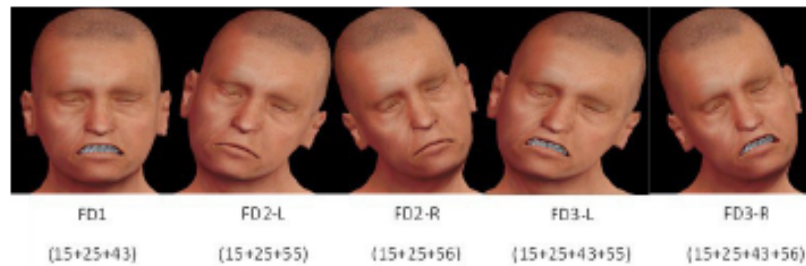


Figure 2. Five classes along with the combination of Action Units.

3.1.2. Transfer Facial Expressions to Static Real Faces Using the First-Order Motion Model (FOMM)

The face swap, bring the face to life, image animation, and Deepfake generations techniques are applications used to replace the face of one person with the face of another sequence. In 2019, Siarohin [35] and his colleagues presented a computer vision and deep learning model, which is called the First-Order Motion Model (FOMM), to generate video sequences in such a way that the object in the source image is animated according to the motion of the driving video. The ability of this model to learn facial expressions is significant without the need to know any prior information about the specific object to animate. The concept of image animation is to synthesize a video using two main parts. The first one is the source image, and the other one is the driving video. The model is trained on a dataset of images and videos for the objects of the same category (e.g., face, body) by identifying key points on the object and then following them to the motion in the video. Recent applications of CNN have proved to mimic realistic human faces. Training networks on a large number of images and video datasets can generate realistic talking persons. A source image of someone can be animated to the target poses of another one in the driving video [36]. The FOMM combines the appearance extracted from the source image and the motion derived from the driving video. The framework described in [35] has achieved satisfactory results on a variety of object categories. Their model has pre-processed the dataset, extracting an initial bounding box in the first video frame. Then, it tracks the object until it is too far away from the initial position. After that, the video frames use the smallest crop containing all the bounding boxes. This process is repeated until the end of the sequence. Then, it filters out sequences that have a resolution lower than 256×256 , and the remaining videos are resized to 256×256 , preserving the aspect ratio to obtain a more realistic video where the head moves freely within the bounding box. The model uses 19,522 training videos and 525 test videos, with lengths varying from 64 to 1024 frames. This project has adapted the FOMM to capture facial expressions for various images. The model was trained to reconstruct the training videos by combining a single frame and a learned potential characterization of the motion in the driving video. At test time, we applied our model to pairs composed of the source image and each frame of the driving video and perform image animation of the source object. The model was trained and tested with different datasets containing various objects. More precisely, the method automatically produced videos by combining the appearance extracted from a source image with motion patterns derived from a driving video. For instance, a facial image of a certain person could be animated following the facial expressions of another person as shown in the sequence of frames from Figure 3. The method was employed in this proposed study to generate and expand a more realistic dataset for real people's faces by transferring this combination of involuntary facial expressions and head poses of patients under risk of deterioration by animating the facial expression, eyeball movement, and head poses of real faces in a source image based on the motion of a facial expression and head poses of avatars in a driving video as shown in Figure 3. We applied the FOMM using a pre-trained deep learning method, and the raw data consisted of driving videos of 3D animated characters displaying the specific expressions of a patient in deterioration

APPENDIX A. PUBLISHED PAPER

and real human faces from an open database known as the Celebrity Face Image Dataset. These expressions were faithfully transferred to the various source images of different real people's faces as shown in Figures 3 and 4. The model was implemented by utilizing source images of real human faces from an open database known as the Celebrity Face Image Dataset, and the results of facial expressions, head poses, eyeball movement, and other actions from the videos transferred to the source images were considered of good quality and realistic. Transferring the facial expressions data to static real faces and bringing them to life was an essential task to train the model on the facial expressions of real human faces. Creating a realistic robust model that can be applied in the real world on real faces can be achieved by utilizing a model that can bring the real faces of facial images to life as can be seen in Figures 3 and 4.

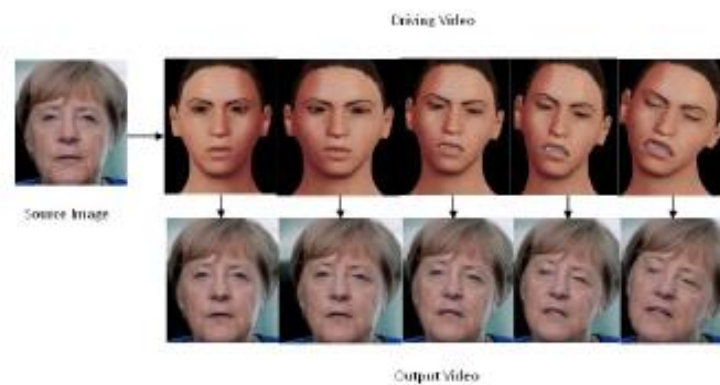


Figure 3. Frames of video sample after utilizing FOMM to transfer facial expressions from avatars to real facial images.

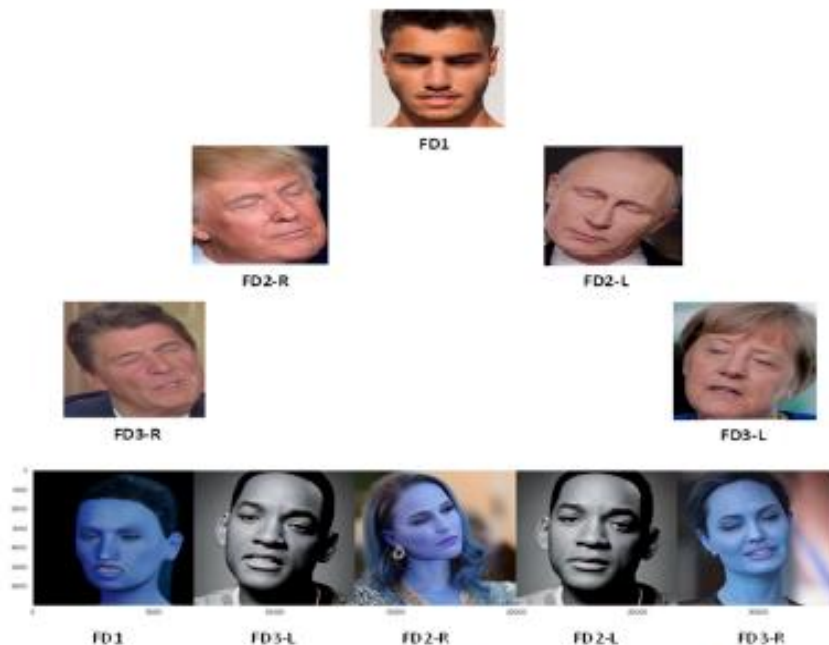


Figure 4. Samples of five classes of facial frames representing five classes.

APPENDIX A. PUBLISHED PAPER

In summary, after generating and recording the facial expressions of each avatar, the facial expressions of avatars were transferred to static images of real people's faces using the Celebrity Face Image Dataset that is available as an open database on Kaggle. Using the FOMM, the number of generated videos was expanded to reach 176 video clips that had colourful sequences of frames with a framerate of around 25 fps and a length of each video (11–12) seconds, so we had (275–300) frames in each video. The paper presents 176 generated videos, and there were around 50,550 frames for all videos. The five facial expressions are shown in Figure 4.

3.2. Pre-Processing Dataset

The pre-processing methods have a great impact on improving the performance of the learning process and model generalization by enhancing the quality of the dataset, minimizing noise, introducing variability, and providing standardized input data for machine learning models. Choosing the appropriate methods is based on the nature of the data, characteristics of data, requirements of the machine learning model, and the target task. It is crucial to achieve a balance between increasing the variability and preserving the essential features of the dataset. This section produces pre-processing methods that have been employed for the proposed system. To achieve consistency across the various image datasets, it was crucial to adjust the dimensions of images by resizing them to a particular size.

3.2.1. Face Detection Technique

Detecting and identifying faces is considered a crucial step in FER for several reasons. Its significant rule lies in providing relative data by focusing on the face region that contains the essential features and patterns to capture detailed information on facial expressions and classify their types by involving subsequent analysis. The faces have a wealth of information expressed by facial expressions, including the positions, intensity, and appearance of AUs for specific facial muscle movements. Therefore, providing the model with the region of interest by localizing and aligning the face area aids in isolating the face from the background and introducing the relevant features that result in reducing the impact of introducing irrelevant data, like backgrounds with noise that may affect the accuracy of model prediction, along with unnecessary computations which reduce the model performance. The proposed framework used the open-source programming packages called Mediapipe (version 0.10.9) for face detection. Mediapipe is considered a well-known method for face detection and facial landmark location, and it can be used for the recognition of facial features and expressions through single facial images or a continuous stream of facial images. Figure 5 shows how the Mediapipe provides a pre-trained face mesh model that can detect face and facial landmarks such as the eyes, nose, mouth, eyebrows, jawline, etc., through facial frames.

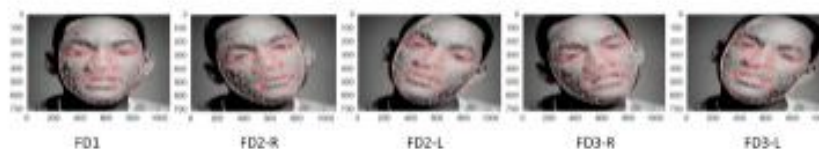


Figure 5. Facial frames samples for each class after pre-processing using face mesh as a face detection technique.

The architecture of the pre-trained model for locating the face and its landmarks was based on a combination of methods, including computer vision and deep-learning algorithms, that were trained to localize facial landmarks in images and frames. The convolution neural network (CNN) is the deep learning method for detecting faces and localising their facial landmarks. Its architecture consists of multiple convolution layers followed by

APPENDIX A. PUBLISHED PAPER

the pooling layers and fully connected layers. This model was trained to automatically learn the hierarchical facial features of images to achieve an accurate prediction.

Each generated video was labelled and categorized into one class of five classes according to its Face Display (FD). Figure 6 illustrates the distribution and number of samples for each class.

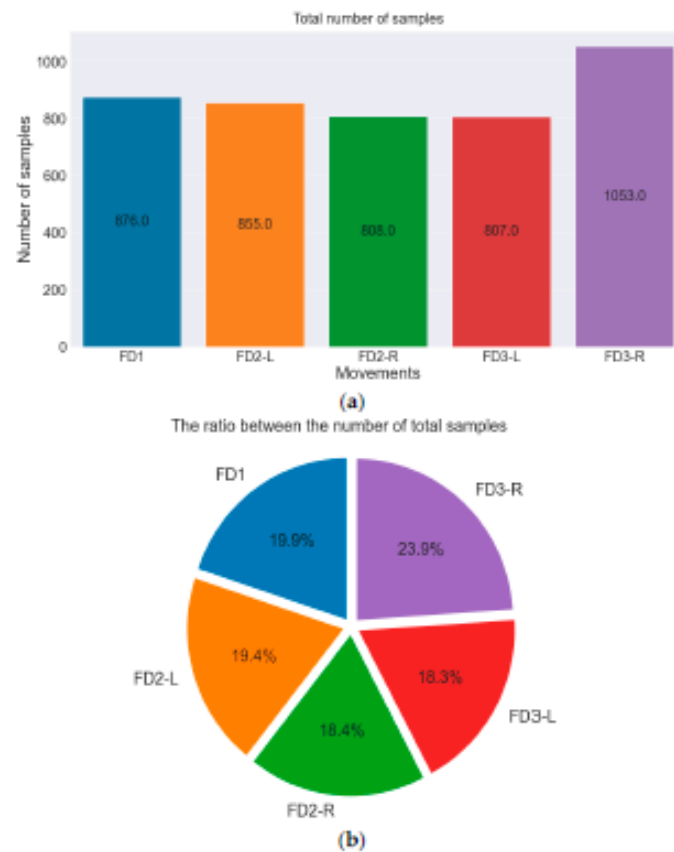


Figure 6. Number and ratio of samples in each class for the whole dataset. (a) The total number of samples is represented by column chart. (b) The ratio of samples in each class.

The whole dataset was then split into training and test datasets. The test data were essential to evaluate model performance on unseen data. The split was performed at 15% for test data and 85% for training data, as illustrated in Figure 7. It is worth noting that any proposed model performs more precisely when it is fed with a rich, sufficient, and diverse dataset.

APPENDIX A. PUBLISHED PAPER

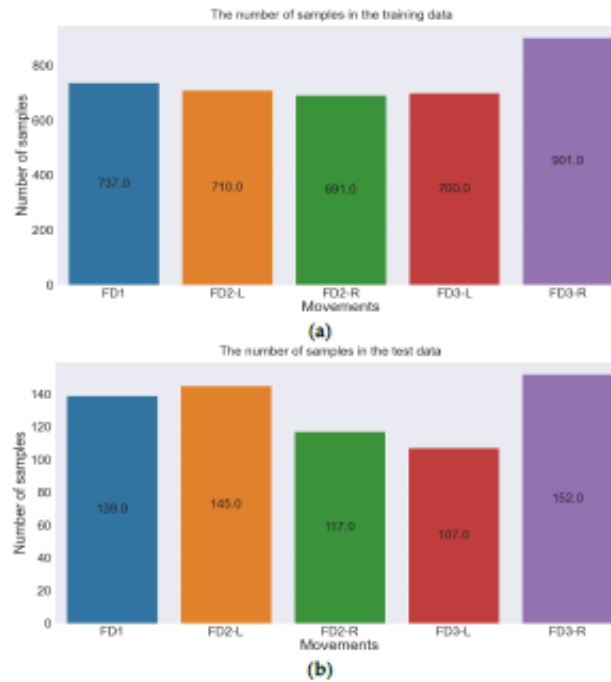


Figure 7. Number of samples in training and test dataset. (a) The number of samples in the training dataset. (b) The number of samples in the test dataset.

3.2.2. Oversampling

The final step in the pre-processing method is the oversampling method, which is considered an effective process in machine learning to handle imbalanced classes and improve model performance by training it with a balanced training set. The oversampling method was only applied to training datasets to avoid data leakage. Again, evaluating the model performance on an imbalanced test dataset is crucial to assess its ability to perform real-world generalization. Figure 8 shows the training dataset before and after applying the oversampling method.

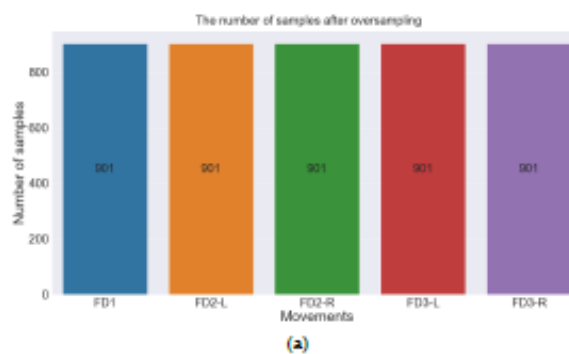


Figure 8. Cont.

APPENDIX A. PUBLISHED PAPER

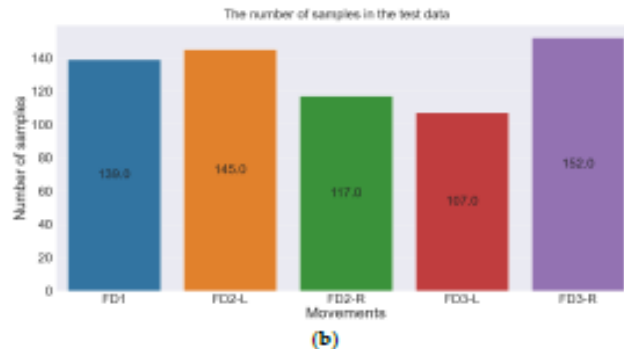


Figure 8. Number of training samples before and after oversampling method. (a) Number of samples of training dataset before oversampling. (b) Number of samples of training dataset after oversampling.

3.3. Proposed Convolution Long Short-Term Memory (ConvLSTM) Model

LSTM can handle temporal input data and achieve high accuracy of prediction; however, it suffers from capturing spatial data, resulting in failing to capture features of spatial data. Therefore, Xingjian and other contributors [37] addressed this problem and developed ConvLSTM, which replaces the state-to-state transition operations in the LSTM with convolution operations. It involves convolution operation within the LSTM structure, and it is particularly popular in computer vision and video analysis tasks as it has demonstrated remarkable success in capturing and handling complex dynamic patterns within image sequences and video streams. The ConvLSTM model expands the traditional LSTM capabilities by involving convolution layers to propose a method that allows the model to learn and retain spatial dependencies in the input sequential data producing an effective prediction model for tasks involving sequential data with spatial characteristics such as video analysis, spatiotemporal modelling, and image sequence processing [38].

3.3.1. Convolution Layers

These layers are responsible for performing convolution operations on input data to capture spatial patterns and relationships to extract relative features [37,38].

3.3.2. LSTM Cells

These cells are involved in capturing temporal dependencies in the input sequential data [37,38]. Each cell includes three types of gates: the input gate, forget gate, and output gate. These gates are responsible for regulating the flow of information through the cell, allowing the network to retain or discard information over time [39]. By combining convolution layers and LSTM cells, the model can effectively process both spatial and temporal dependencies in the sequential data. Therefore, it is considered suitable for handling tasks such as video prediction, action recognition, facial emotion recognition, and other tasks where understanding and analysing both the spatial and temporal aspects of data is crucial. Therefore, this paper proposed the ConvLSTM model to recognize facial expressions through frames of video stream due to its ability to capture both spatial and temporal dependencies in facial expressions over time. Figure 9 shows the ConvLSTM structure [40] where the new memory C_t and output H_t will be generated by updating the internal memory C_{t-1} to the current input X_t and the previous output H_{t-1} .

APPENDIX A. PUBLISHED PAPER

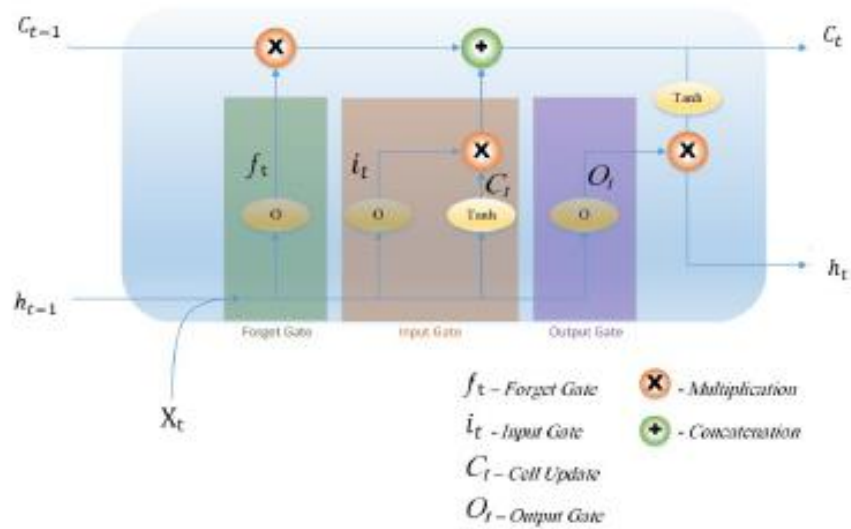


Figure 9. Structure of ConvLSTM [40].

The mathematical expression of the ConvLSTM in the updated gates is given as follows:

$$f_t = (W_x f * X_t + W_h f * h(t-1) + W_c f * C(t-1) + b_f) \quad (1)$$

$$i_t = (W_x i * X_t + W_h i * h(t-1) + W_c i * C(t-1) + b_i) \quad (2)$$

$$O_t = (W_x o * X_t + W_h o * h(t-1) + W_c o * C_t + b_o) \quad (3)$$

$$C_t = f_t C(t-1) + i_t \tanh(W_x c * X_t + W_h c * h(t-1) + b_c) \quad (4)$$

$$h_t = O_t \tanh(C_t) \quad (5)$$

where $*$ refers to convolution operation, and \otimes refers to the Hadamard product. W_{gf} , W_{gi} , and W_{co} refer to the weight matrices.

All the weight matrices and bias vectors will be updated in each update process.

In this model, a background removal procedure was applied before the generation of the extraction vector to avoid dealing with multiple problems that may occur such as noise of the background, distance from the camera, light, irrelevant data, etc. Then, an expressional vector has been applied to detect and characterize the 5 various kinds of patients' faces under deterioration. It was possible to correctly highlight the class label of facial expression with 99.4% accuracy. The proposed system phases are depicted in Figure 10.

The model was trained and evaluated using the k-fold cross-validation process, which helped ensure that the model generalised well to unseen data. Instead of relying on a single train-test split, which might lead to overfitting or underfitting, k-fold cross-validation trained the model on multiple different subsets of the data, which resulted in better learning the underlying patterns and avoiding overfitting to any particular subset.

APPENDIX A. PUBLISHED PAPER

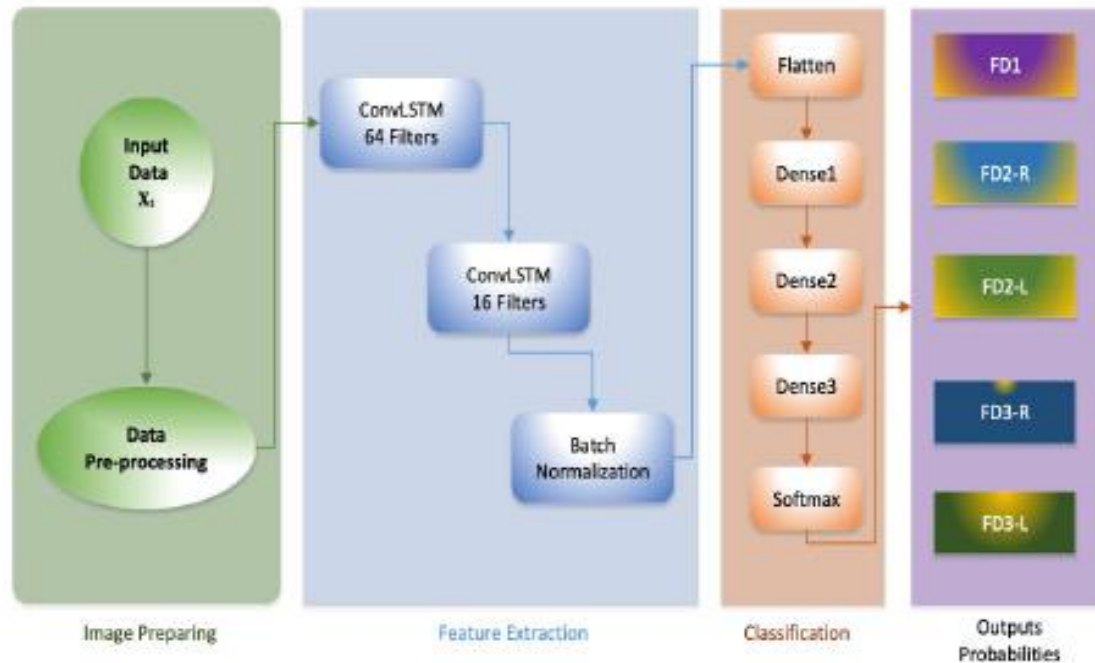


Figure 10. The proposed model architecture.

4. Results and Limitations

The proposed model has been trained and tested on the generated dataset. The dataset includes video frames for five classes of specific facial expressions (FD1, FD2-L, FD2-R, FD3-L, FD3-R) for various facial landmarks, skin tones, and ethnicities representing patients at risk of deterioration. Achieving optimal results depends on many factors, such as the quality, quantity and diversity of the dataset, the effectiveness of feature extraction methods, the model structure, experimentations, and fine-tuning. After the data training stage, the model has to be evaluated for its reliability by testing their ability to handle and master the target task. The evaluation of machine learning models is based on essential metrics such as accuracy. The target of a machine learning engineer or designer is to achieve the highest model accuracy, and this measurement represents the model's ability to find the features and relationships between data that relate to the target task. The accuracy is focused on the number of true predicted samples and calculated by finding the number of correctly predicted samples to the overall number of predictions. There are four essential measures used for estimating model performance, including: 1. True positives (TP): the number of correctly predicted samples. 2. True negatives (TN): the number of rightly predicted values as negative. 3. False positives (FP): the number of positive samples that are wrongly predicted. 4. False negatives (FN): the number of negative samples that are incorrectly predicted. The correct predictions of the model include the true positives and the true negatives, while the model misleading includes the false negatives and false positives. The accuracy of the model can be calculated by the following formula [41,42]:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True negatives} + \text{False positives} + \text{False Negatives}} \quad (6)$$

The accuracy metric is a straightforward measurement; however, it cannot be considered a sufficient evaluation for all tasks due to some limitations. For instance, it might be the improper measurement in evaluating imbalanced classes where there is a substan-

APPENDIX A. PUBLISHED PAPER

tial difference in the number of samples in a class compared with the other classes. It may result in a metric of accuracy being very high because of its correct prediction of the majority class, even if the model performs poorly in the other minority classes. Another metric is precision, which is responsible for measuring the ability to capture the number of correctly predicted samples of positive class. It can be calculated by finding the ratio of correct sample predictions to the overall number of samples identified as positive class. The proportion between true positives that are correctly identified and the total of both true positives and false positives can be calculated in the following formula [43]:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

Recall is an evaluation metric that is sometimes known as sensitivity, especially in the medical and biological fields, or true-positive rates due to its ability to provide an accurate evolution of model performance in identifying the positive samples. It records the ability to identify positive samples and can be measured by finding the ratio between the true positives and the total number of positive samples [44].

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{8}$$

Another popular metric for evaluating model performance is the F1 score. It is considered the harmonic mean of precision and recall, providing a balance between them and serving as an effective metric in imbalanced classes. Its importance lies in evaluating a model's ability to detect true positives and false negatives. The equation for calculating the F1 Score is as follows [45]:

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

Table 3 illustrates these metrics, which provide an evaluation of the proposed ConvLSTM model. The evaluation of model performance is measured by testing the prediction of the model on unseen or new data during the testing process, recognizing relevant features in unseen new data. One of the most common evaluation methods is the confusion matrix which uses four essential components: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), in assessing and evaluating the performance of classification models. Figures 11 and 12 show the accuracy and loss of the testing dataset for the proposed model.

Figure 13 presents the confusion matrix that summarises and visualizes the performance of the proposed model. Each row of the matrix presents facial expressions in the actual class, while each column represents facial expressions in the predicted class.

Figures 14 and 15 illustrate all evaluated measurements of the proposed model.

The five evaluation measurements for five classes are illustrated in Table 3.

Table 3. Evaluation metrics for each class: the total mean performance.

Class Name	Facial Expression	Precision	Recall	F1 Score	Accuracy
FD1	AU (15 + 25 + 43)	100%	100%	100%	100%
FD2-R	AU (15 + 43 + 55)	100%	100%	100%	100%
FD2-L	AU (15 + 43 + 56)	99%	100%	100%	99%
FD3-R	AU (15 + 25 + 43 + 55)	100%	100%	100%	100%
FD3-L	AU (15 + 25 + 43 + 56)	100%	99%	100%	100%
	Mean	99.8%	99.8%	100%	99.8%

The above measurements provide insights into various aspects of model performance. The precision, recall, F1 score, and accuracy recorded 99.8%, 99.8%, 100%, and 99.8%,

APPENDIX A. PUBLISHED PAPER

respectively. This study shows very promising outcomes in detecting the deterioration of patients from their facial expressions. However, the limitation of this project is that real-life data samples could not be collected due to ethical procedures as the data related to patients in critical care units and intensive care units. The generated data are based on the psychologists' study presented by [3], which helped to introduce avatars mimicking the exact specific five categorical facial expressions that show patients suffering from deterioration.

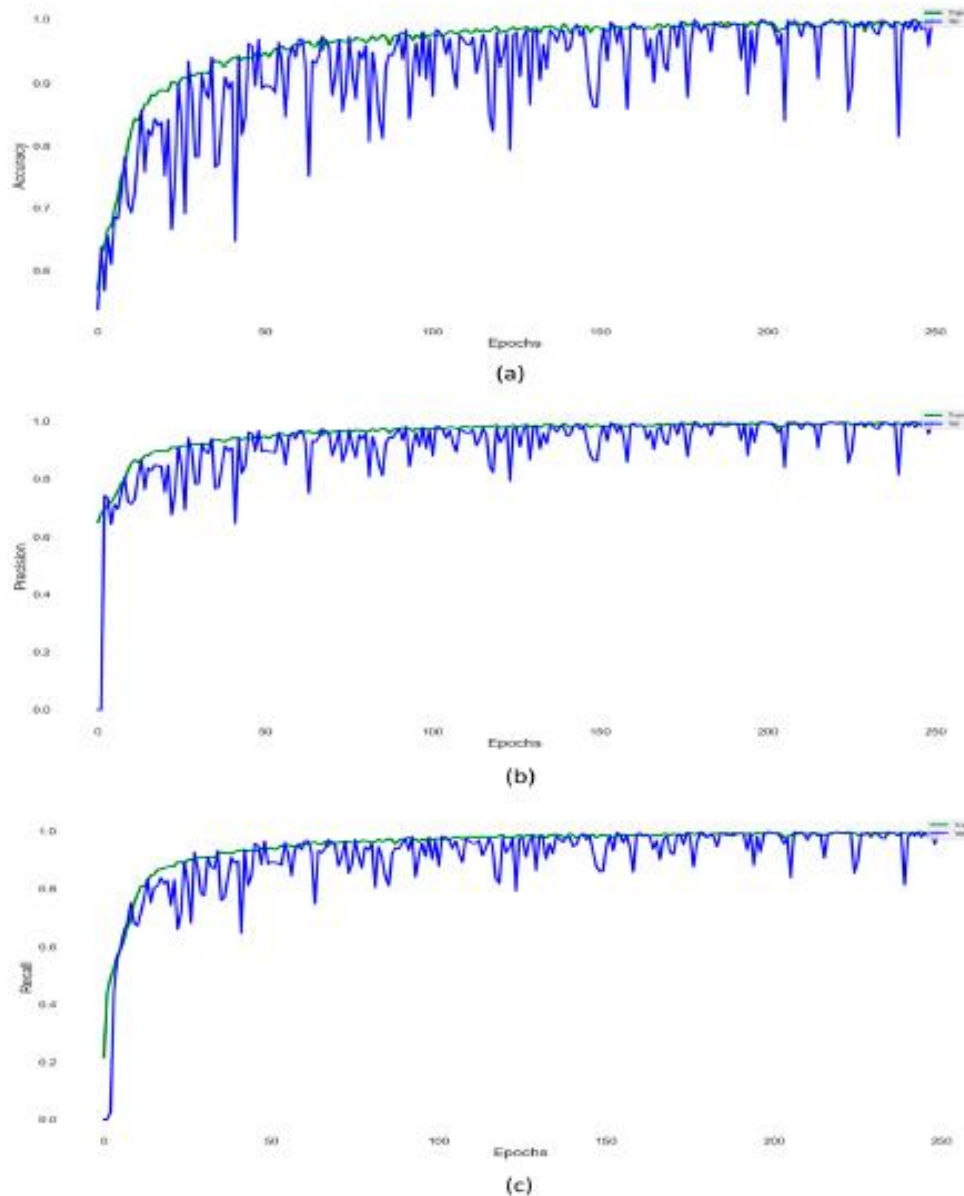


Figure 11. (a) Evaluation metrics of model performance. Accuracy of the proposed model (b) Precision of the proposed model. (c) Recall of the proposed model.

APPENDIX A. PUBLISHED PAPER

We conducted experiments with other deep learning models, such as Vision Transformers, on the generated dataset. However, the results were not satisfactory due to the spatial-temporal nature of the dataset's features. In contrast, the CNN model has a significant ability to explore and recognise spatial features, while the LSTM model is well known for its capability of capturing temporal features. Consequently, the ConvLSTM model achieved state-of-the-art results in predicting the facial expressions (FEs) of patients at risk of deterioration.

Figure 16a shows the ROC curve and Figure 16b shows the precision-recall (PR) curve, which are used to evaluate the performance of a classifier, especially when dealing with imbalanced datasets.

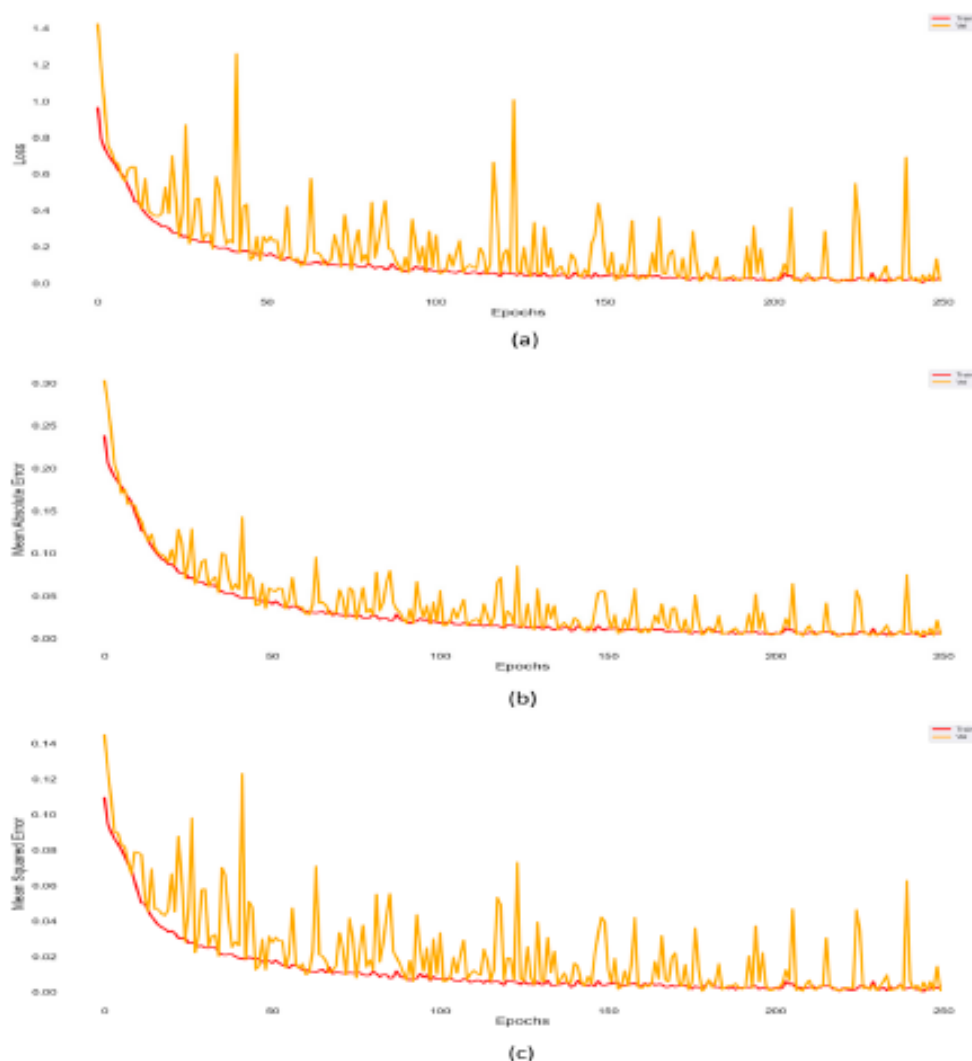


Figure 12. Loss, Mean Square Error and Mean Absolute Error: (a) Loss of the predicted model. (b) Mean Square Error of the predicted model. (c) Mean Absolute Error.

APPENDIX A. PUBLISHED PAPER

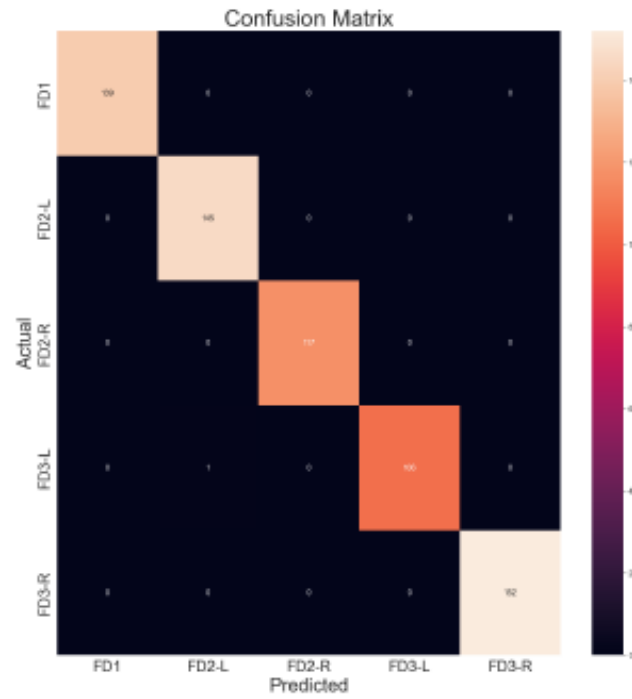


Figure 13. Confusion matrix.

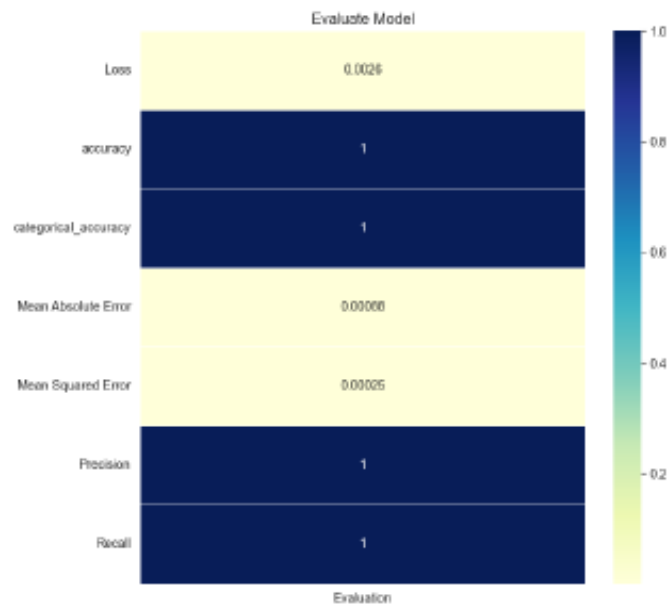


Figure 14. Evaluation of the model.

APPENDIX A. PUBLISHED PAPER

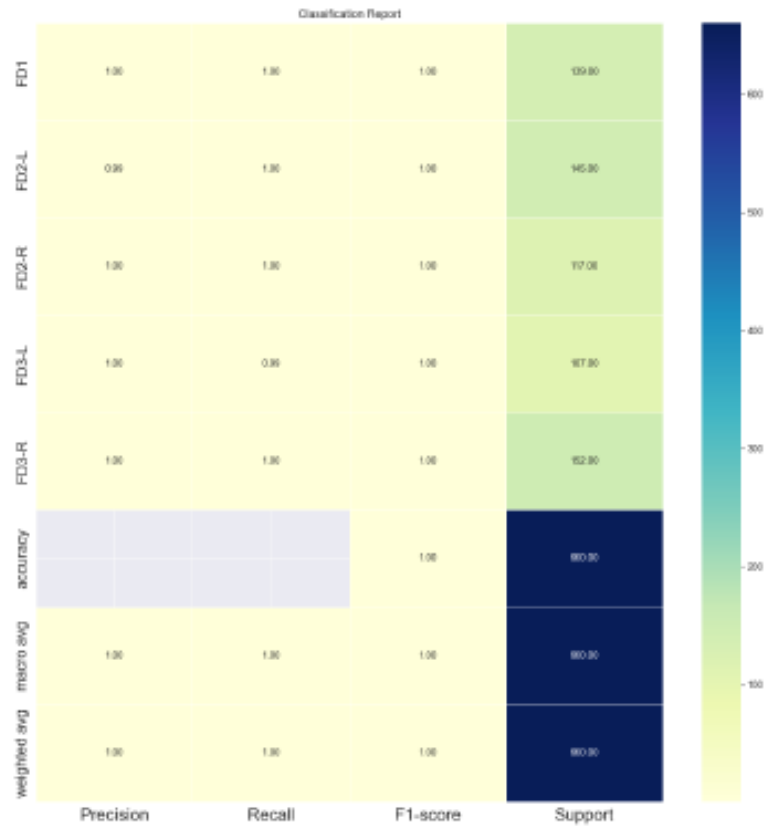
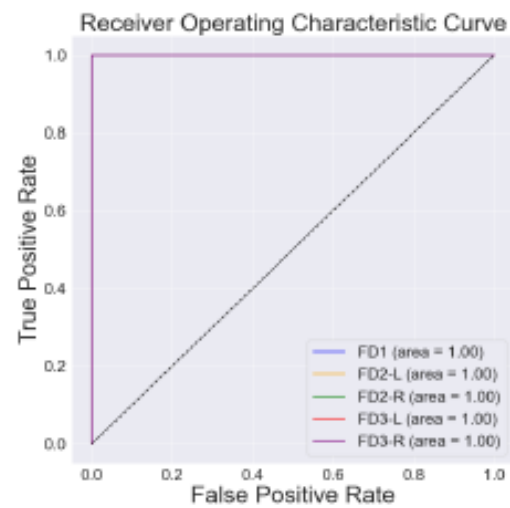


Figure 15. Classification report.



(a)

Figure 16. Cont.

APPENDIX A. PUBLISHED PAPER

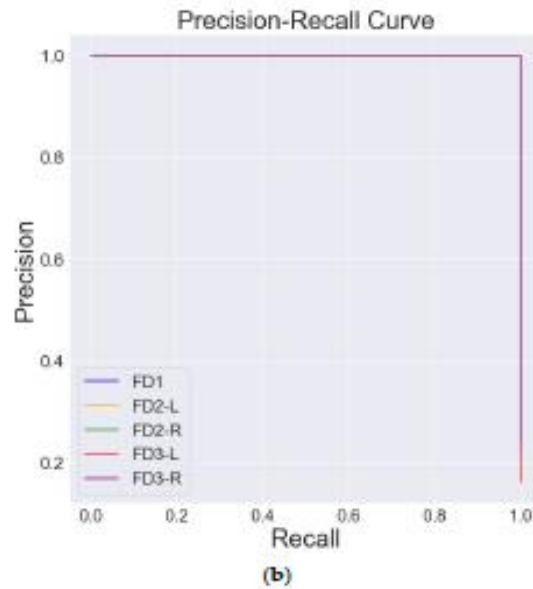


Figure 16. Evaluating model by Receiver Operating Characteristics Curve (ROC) Precision-Recall Curve. (a) ROC. (b) Precision-Recall Curve.

In addition, the model has been evaluated using the unseen data separated from the whole dataset before training, and the model also shows high predicted results as shown in Figure 17.

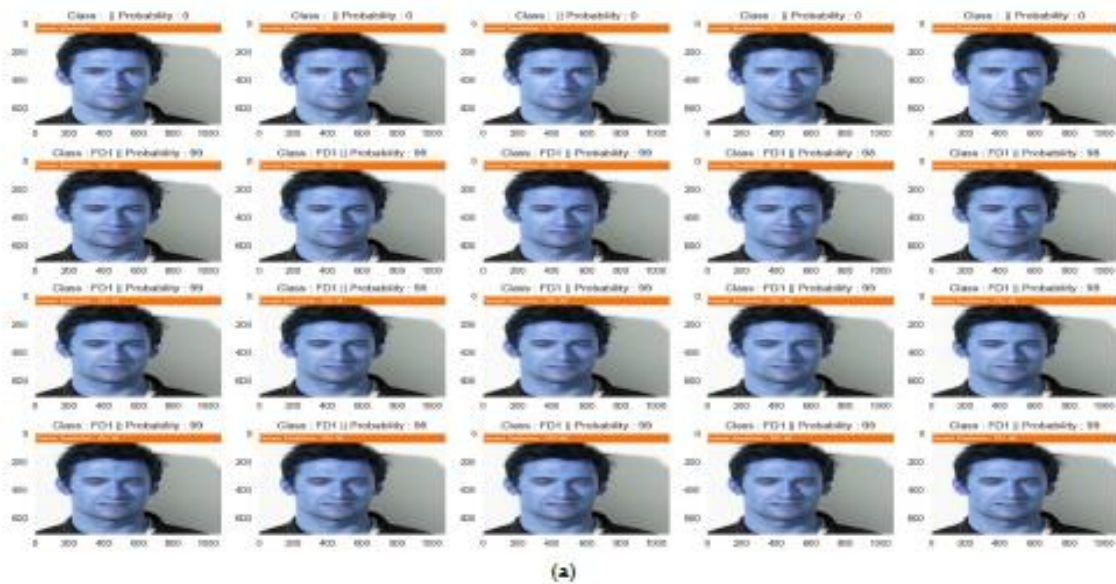


Figure 17. Cont.

APPENDIX A. PUBLISHED PAPER



Figure 17. Cont.

APPENDIX A. PUBLISHED PAPER

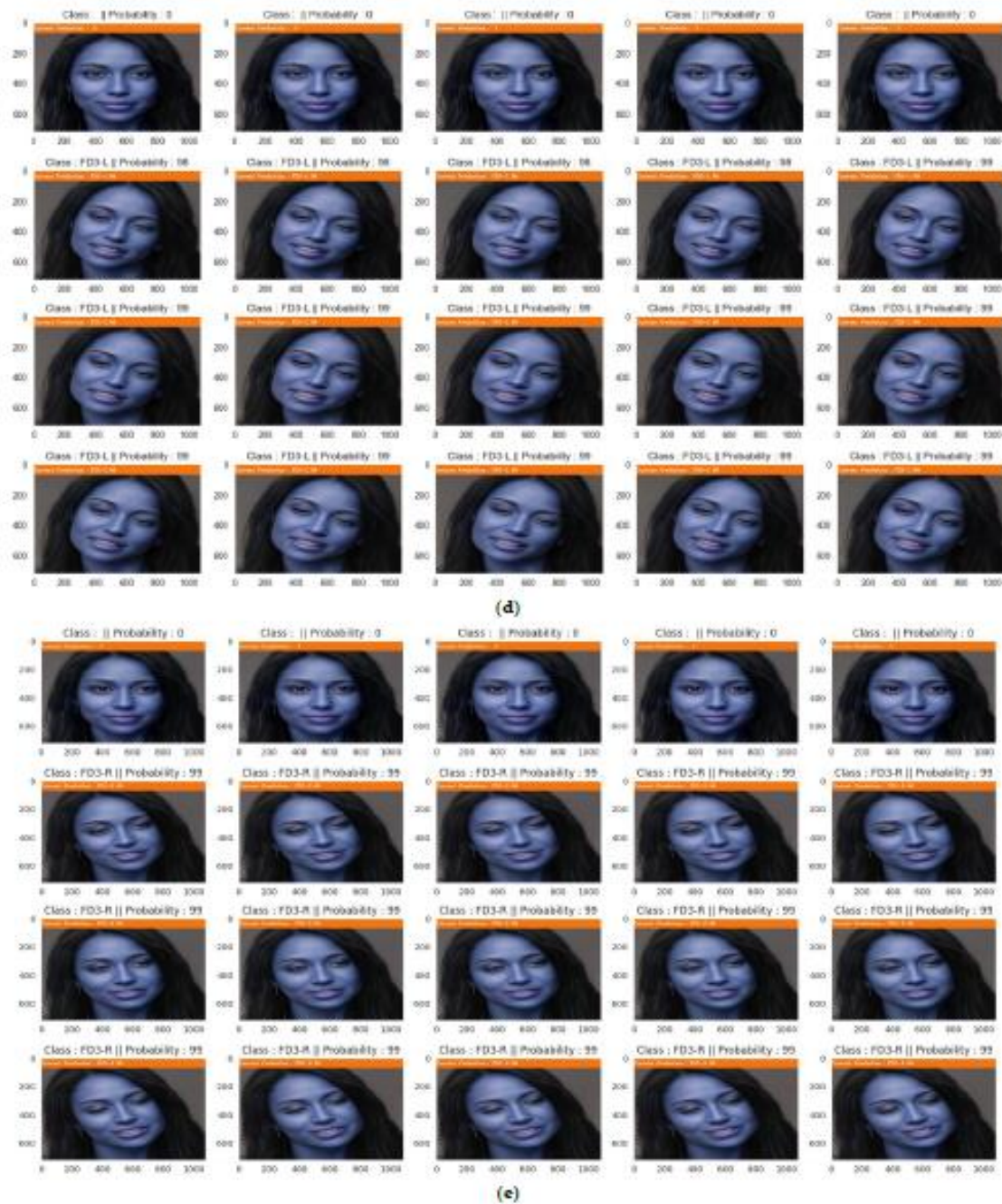


Figure 17. The percentage of accuracy of model prediction for unseen data for different classes. (a) Accuracy of prediction of unseen data predicted as Class FD1. (b) Accuracy of prediction of unseen data predicted as Class FD2-L. (c) Accuracy of prediction of unseen data predicted as Class FD2-R. (d) Accuracy of prediction of unseen data predicted as Class FD3-L. (e) Accuracy of prediction of unseen data predicted as Class FD3-R.

APPENDIX A. PUBLISHED PAPER

5. Conclusions

Our faces hold and show valuable clues about human emotions and their intentions [46]. FER has been intensively studied for the last few decades in computer vision, due to its importance to improving communication with individuals and generate empathetic responses. Early detection of signs of patients' deterioration from facial expressions is a challenging task for healthcare professionals. Therefore, this paper has concentrated on proposing suitable methods, employing deep learning algorithms as a solution for identifying signs of patients' deterioration through their FEs. With recent technologies and advancement in computer vision, pattern recognition, and machine learning, it is possible to detect and characterize FEs through images and video streams with high accuracy using DNN models such as ConvLSTM model. The main objective of this research was to design a framework for automatic FER to predict FEs of patients at risk of deterioration. The proposed system used a generated database called PRD-FE comprising five different combination sets of AUs (FD1, FD2-R, FD2-L, FD3-R, and FD3-L), representing FEs of deterioration risk. This paper presents a framework for automatic FER based on facial landmarks and ConvLSTM architecture, achieving state-of-the-art results with an accuracy of 99.89%. The proposed system has used a generated database that includes five classes of patients under deterioration, i.e., FD1, FD2-R, FD2-L, FD3-R, and FD3-L. Employing the facial landmarks detection technique resulted in improving the prediction of the proposed model, achieving a significant accuracy of around 99.8%. Future work will concentrate on collecting real-world data to further validate the proposed models and present them as integrated systems with other medical assessment systems to enhance the chances of human survival.

Author Contributions: Conceptualization, Z.A.-T. and M.A.R.; methodology, Z.A.-T. and M.A.R.; software, Z.A.-T. and M.A.R.; validation, Z.A.-T. and M.A.R.; investigation, J.M.-C. and M.L.M.G.; resources, J.M.-C. and M.L.M.G.; data curation, Z.A.-T. and M.A.R.; writing—original draft preparation, Z.A.-T. and M.A.R.; writing—review and editing, Z.A.-T. and M.A.R.; visualization, M.A.R.; supervision, M.A.R.; project administration, M.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sheffield Hallam University, grant number B3036983.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manalu, H.V.; Rifai, A.P. Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intell. Syst. Appl.* **2024**, *21*, 200339. [\[CrossRef\]](#)
2. Cuesta, J.M.; Singer, M. The stress response and critical illness: A review. *Crit. Care Med.* **2012**, *40*, 3283–3289. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Madrigal-Garcia, M.I.; Rodrigues, M.; Shenfield, A.; Singer, M.; Moreno-Cuesta, J. What faces reveal: A novel method to identify patients at risk of deterioration using facial expressions. *Crit. Care Med.* **2018**, *46*, 1057–1062. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Street, R.L., Jr.; Makoul, G.; Arora, N.K.; Epstein, R.M. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Primt Educ. Couns.* **2009**, *74*, 295–301. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Jones, D.; Mitchell, L.; Hillman, K.; Story, D. Defining clinical deterioration. *Resuscitation* **2013**, *84*, 1029–1034. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Alasad, J.; Ahmad, M. Communication with critically ill patients. *J. Adv. Nurs.* **2005**, *50*, 356–362. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Ye, C.; Wang, O.; Liu, M.; Zheng, L.; Xia, M.; Hao, S.; Jin, B.; Jin, H.; Zhu, C.; Huang, C.J.; et al. A real-time early warning system for monitoring inpatient mortality risk: Prospective study using electronic medical record data. *J. Med. Internet Res.* **2019**, *21*, e13719. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Herr, K.; Coyne, P.J.; Ely, E.; Gelinias, C.; Manworzen, R.C. Pain assessment in the patient unable to self-report: Clinical practice recommendations in support of the ASPMN 2019 position statement. *Prim Manag. Nurs.* **2019**, *20*, 404–417. [\[CrossRef\]](#)
9. Odell, M.; Victor, C.; Oliver, D. Nurses' role in detecting deterioration in ward patients: Systematic literature review. *J. Adv. Nurs.* **2009**, *65*, 1992–2006. [\[CrossRef\]](#)

APPENDIX A. PUBLISHED PAPER

10. Guo, X.; Zhang, Y.; Lu, S.; Lu, Z. Facial expression recognition: A review. *Multimed. Tools Appl.* **2023**, *83*, 23689–23735. [\[CrossRef\]](#)
11. Prakash, M.; Ravichandran, T. An efficient resource selection and binding model for job scheduling in grid. *Eur. J. Sci. Res.* **2012**, *81*, 450–458.
12. Mehrabian, A. *Nonverbal Communication*; Routledge: London, UK, 2017.
13. Ekman, P.; Friesen, W.V. *Facial Action Coding Systems*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
14. Rudovic, O.; Pavlovic, V.; Pantic, M. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 944–958. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Cascella, M.; Schiavo, D.; Cuomo, A.; Ottaiano, A.; Perri, F.; Patrone, R.; Migliarelli, S.; Bignami, E.G.; Vittori, A.; Cutugno, E.; et al. Artificial intelligence for automatic pain assessment: Research methods and perspectives. *Pain Res. Manag.* **2023**, *2023*, 6018736. [\[CrossRef\]](#)
16. Nagireddi, J.N.; Vyas, A.K.; Sanapati, M.R.; Soir, A.; Manchikanti, L. The analysis of pain research through the lens of artificial intelligence and machine learning. *Pain Physician* **2022**, *25*, E211.
17. Hardas, B.M.; Pokle, S.B. Optimization of peak to average power reduction in OFDM. *J. Commun. Technol. Electron.* **2017**, *62*, 1388–1395. [\[CrossRef\]](#)
18. Rodriguez, P.; Cucurull, G.; Gonzalez, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, E.X. Deep pair: Exploiting long short-term memory networks for facial expression classification. *IEEE Trans. Cybern.* **2017**, *52*, 3314–3324. [\[CrossRef\]](#)
19. Jaswanth, K.; David, D.S. A novel based 3D facial expression detection using recurrent neural network. In Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 3–4 July 2020; pp. 1–6.
20. Sato, W.; Hyniewska, S.; Minemoto, K.; Yoshikawa, S. Facial expressions of basic emotions in Japanese laypeople. *Front. Psychol.* **2019**, *10*, 259. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Gosselin, P.; Kirouac, G.; Doré, E.Y. Components and recognition of facial expression in the communication of emotion by actors. *J. Personal. Soc. Psychol.* **1995**, *68*, 83. [\[CrossRef\]](#)
23. Scherer, K.R.; Ellgring, H. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion* **2007**, *7*, 113. [\[CrossRef\]](#)
24. Lucy, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–25 March 2011; pp. 57–64.
25. Prkachin, K.M.; Solomon, P.E. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* **2008**, *139*, 267–274. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ekman, P.; Friesen, W.V.; Hager, J. Facial action coding system: Research Nexus. In *Network Research Information*; Research Nexus: Salt Lake City, UT, USA, 2002.
27. Gross, J.; Cuesta, J.; Crawford, S.; Devaney, M.; Madrigal-Garcia, M. The face of illness: Analysing facial expressions in critical illness in conjunction with the facial action coding system (FACS). In *Proceedings of the Interist de Care Medicine*; Springer: New York, NY, USA, 2013; Volume 39, p. S265.
28. Chen, J.; Lv, Y.; Xu, R.; Xu, C. Automatic social signal analysis: Facial expression recognition using difference convolution neural network. *J. Parallel Distrib. Comput.* **2019**, *131*, 97–102. [\[CrossRef\]](#)
29. Gunes, H.; Hung, H. Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block. *Image Vis. Comput.* **2016**, *55*, 6–8. [\[CrossRef\]](#)
30. Jaiswal, S.; Valstar, M. Deep learning the dynamic appearance and shape of facial action units. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–8.
31. Sang, D.V.; Van Dat, N. Facial expression recognition using deep convolutional neural networks. In Proceedings of the 2017 9th International Conference on Knowledge and Systems Engineering (KSE), Hue, Vietnam, 19–21 October 2017; pp. 130–135.
32. Chen, X.; Yang, X.; Wang, M.; Zou, J. Convolution neural network for automatic facial expression recognition. In Proceedings of the 2017 International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 13–17 May 2017; pp. 814–817.
33. Al Taei, E.J.; Jasim, Q.M. Blurred Facial Expression Recognition System by Using Convolution Neural Network. *Webology* **2020**, *17*, 804–816. [\[CrossRef\]](#)
34. Mohan, K.; Seal, A.; Krejcar, O.; Yazidi, A. Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 5003512. [\[CrossRef\]](#)
35. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First order motion model for image animation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7137–7147.
36. Malik, Y.S.; Sabahat, N.; Moazzam, M.O. Image animations on driving videos with DeepFakes and detecting DeepFakes generated animations. In Proceedings of the 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 5–7 November 2020; pp. 1–6.
37. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.

APPENDIX A. PUBLISHED PAPER

38. Singh, R.; Saurav, S.; Kumar, T.; Saini, R.; Vohra, A.; Singh, S. Facial expression recognition in videos using hybrid CNN & ConvLSTM. *Int. J. Inf. Technol.* **2023**, *15*, 1819–1830. [[PubMed](#)]
39. Tian, Y.; Zhang, K.; Li, J.; Lin, X.; Yang, B. LSTM-based traffic flow prediction with missing data. *Neurocomputing* **2018**, *318*, 297–305. [[CrossRef](#)]
40. Zhang, L.; Zhu, G.; Mei, L.; Shen, P.; Shah, S.A.A.; Bennamoun, M. Attention in convolutional LSTM for gesture recognition. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1953–1962.
41. Ikram, S.T.; Cherukuri, A.K. Improving accuracy of intrusion detection model using PCA and optimized SVM. *J. Comput. Inf. Technol.* **2016**, *24*, 133–148. [[CrossRef](#)]
42. Thaseen, I.S.; Kumar, C.A. Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *J. King Saud Univ.-Comput. Inf. Sci.* **2017**, *29*, 462–472.
43. Abo-Tabik, M.A. Using Deep Learning Predictions of Smokers' Behaviour to Develop a Smart Smoking-Cessation App. Ph.D. Thesis, Manchester Metropolitan University, Manchester, UK, 2021.
44. Chakravarthi, B.R.; Priyadharshini, R.; Muralidaran, V.; Suryawanshi, S.; Jose, N.; Sherly, E.; McCrae, J.P. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, Hyderabad, India, 16–20 December 2020; pp. 21–24.
45. Lachgar, M.; Hrimech, H.; Kartit, A. Optimization techniques in deep convolutional neuronal networks applied to olive diseases classification. *Artif. Intell. Agric.* **2022**, *6*, 77–89.
46. Arul Vinayakam Rajasimman, M.; Manoharan, R.K.; Subramani, N.; Aridoss, M.; Galey, M.G. Robust facial expression recognition using an evolutionary algorithm with a deep learning model. *Appl. Sci.* **2022**, *13*, 468. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.