Facial Expression Recognition and Intensity Estimation

by

Olufisayo Sunday Ekundayo
218085734

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in the School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal

Durban South Africa, 2022

Supervisor: Prof. Serestina Viriri

Declaration of Authorship

I, Olufisayo Sunday EKUNDAYO, declare that this thesis titled, "Facial Expression Recognition and Intensity Estimation” and the work presented in it are my own. I declare that:

1. The research reported in this thesis, except where otherwise indicated or acknowledged, is my original work;
2. This thesis has not been submitted in full or in part for any degree or examination to any other university;
3. This thesis does not contain other persons‘ data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;
4. This thesis does not contain other persons‘ writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   (a) their words have been re-written but the general information attributed to them has been referenced;

   (b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced;
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Candidate: Olufisayo Sunday Ekundayo

Signature:

Date: **25/02/2022**

As the candidate’s supervisor I approve the submission of this thesis for examination.

Supervisor: Prof. Serestina VIRIRI

Signature:

Date: 25/02/2022

Abstract

Facial Expression is one of the profound non-verbal channels through which human emotion state is inferred from the deformation or movement of face components when facial muscles are activated. Facial Expression Recognition (FER) is one of the relevant research fields in Computer Vision (CV) and Human-Computer Interaction (HCI). Its application is not limited to: robotics, game, medical, education, security and marketing. FER consists of a wealth of information. Categorising the information into primary emotion states only limit its performance. This thesis considers investigating an approach that simultaneously predicts the emotional state of facial expression images and the corresponding degree of intensity. The task also extends to resolving FER ambiguous nature and annotation inconsistencies with a label distribution learning method that considers correlation among data. We first proposed a multi-label approach for FER and its intensity estimation using advanced machine learning techniques. According to our findings, this approach has not been considered for emotion and intensity estimation in the field before. The approach used problem transformation to present FER as a multilabel task, such that every facial expression image has unique emotion information alongside the corresponding degree of intensity at which the emotion is displayed. A Convolutional Neural Network (CNN) with a sigmoid function at the final layer is the classifier for the model. The model termed ML-CNN (Multilabel Convolutional Neural Network) successfully achieve concurrent prediction of emotion and intensity estimation. ML-CNN prediction is challenged with overfitting and intraclass and interclass variations. We employ Visual Geometric Graphics-16 (VGG-16) pretrained network to resolve the overfitting challenge and the aggregation of island loss and binary cross-entropy loss to minimise the effect of intraclass and interclass variations. The enhanced ML-CNN model shows promising results and outstanding performance than other standard multilabel algorithms. Finally, we approach data annotation inconsistency and ambiguity in FER data using isomap manifold learning with Graph Convolutional Networks (GCN). The GCN uses the distance along the isomap manifold as the edge weight, which appropriately models the similarity between adjacent nodes for emotion predictions. The proposed method produces a promising result in comparison with the state-of-the-art methods.

Declaration

The work described in the thesis was carried out in the School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal from July 2018 to August 2021. This dissertation was completed under the supervision of Professor Serestina Viriri.

This study represents original work by the author and has not been submitted in any form for any degree or diploma to any other tertiary institution. Where use was made of the work of others it has been duly acknowledged in the text.

Acknowledgements

My First appreciation goes to the Almighty God, the giver of life, knowledge, wisdom, understanding, good health, divine protection and provisions needed to complete the programme.

I sincerely appreciate all the staff members in the School of Computer Science, UKZN, Westville campus, for their support. My profound appreciation goes to my supervisor Prof. Serestina Viriri for his mentorship, brotherly care, patience and friendliness throughout the programme. Your kindness towards me shall not be forgotten, I am grateful Sir.

To all my family members, starting from my mother, Mrs Comfort Ekundayo, I thank you for all your prayers towards my success. I am grateful to you, Pastor and Mrs Osanyinbi Folorunso. I could not imagine this programme without your full support both spiritually and financially. I could proudly and loudly say that this is your dream that finally came to true. Your fatherly role could not be overemphasised, thank you. Mr Osanyinbi Sojibola, I appreciate you also. Mr Abiodun Ekundayo, time will not permit me to state your contributions to the success of this programme both in cash and in kinds despite all odds. I appreciate you, Sir.

To my colleagues and my friends; I appreciate Pastor and Mrs Sule for the role they play as a brother and sister, other colleagues; Dr Adegun Adekanmi, Dr Agbo-Ajala Bosun, Dr Omral Salih, Dr Roland, Mrs Janet and Mr Mustafa Oloko-Oba, thanks for accommodating me into your team. Without any reservation, I appreciate working with you all. Prophet Steve Olayinka Salawu, Mr Folorunsho Olaiya, Mr Olupitan Gabriel, Dr and Mrs Fagbola Tayo, Mr Ajayi Akinjide, Mr Omotayo Olakunle, POP G. and Prophet Bello Samuel, Mrs Kele, I appreciate all your encouragement, Financial support and timely help. Thank you all.

I am indebted to you, Mrs Hellen Olufunmilayo Ekundayo, my lovely wife, I appreciate your patience and your grace to stand in the gap for me. I also appreciate my children, Emmanuel, Daniel and Rachael Ekundayo, for your cooperation. I sincerely understand your feelings of not having your father around.

Dedication

This thesis is dedicated to Almighty God.

Contents

List of Tables

List of Figures

List of Publications

1. **Olufisayo Ekundayo and Serestina Viriri**, "Facial Expression Recognition: A Review of Trends and Techniques", in*IEEE Access,* vol. 9, pp. 136944-136973, **(2021)**, DOI:text 10.1109/ACCESS.2021.3113464.

2. **Olufisayo Ekundayo, Serestina Viriri**, "Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach", *Advance in Visual Computing*, LNCS Springer, vol. 12510, pp. 581-592, **(2020)**. DOI:https://doi.org/10.1007/978-3-030-64559-5_46.

3. **Olufisayo Ekundayo, Serestina Viriri**, "Deep Forest Approach For Facial Expression Recognition", *Image and Video technology*, LNCS Springer, Vol. 11994, pp. 149-161, **(2020)**. DOI:https://doi.org/10.1007/978-3-030-39770-8_12.

4. **Olufisayo Ekundayo, Serestina Viriri**, "Facial Expression Recognition: A Review of Methods, Performances and Limitation", *IEEE International Conference on Information Communication Technology and Society (ICTAS)*, pp. 1-6, **(2019)**. DOI:10.1109/ICTAS.2019.8703619.

5. **Olufisayo Ekundayo and Serestina Viriri**, "Multilabel Convolutional Neural Network for Facial Expression Recognition and Ordinal Intensity Estimation", *PeerJ Computer Science* (Accepted for Publication).

6. **Olufisayo Ekundayo, Serestina Viriri**, "Facial Expression Recognition Using Manifold learning and Graph Convolutional Network", Manuscript submitted to *International Journal of Intelligent Systems*, (under peer review).

Chapter 1

Introduction

1.1 Facial Expression Recognition

In recent years, digitisation has become the order of the day, manual approaches to getting tasks done gradually fade away and become history due to evolution in computations and the challenges in this present age. Artificial Intelligence (AI), has in some capacities, helped in building intelligent systems. Computer Vision (CV) and Human-Computer Interaction (HCI) contributions to the AI world cannot be overemphasised. Facial Expression Recognition (FER) is a field of computer vision where human affective states are determined from facial deformations. Figure 1.1 shows the relationship of FER with AI. Automation of FER has many applications that are not limited to; security [3] [4], health and medical [5] [6] [7], commerce [8], advertisement [9], and education [10].

The basic emotions inferred from the analysis of FER are categorised into six different classes [11] (Anger, Disgust, Fear, Happy, Sadness, Surprise). Although Martinez and Valstar [12] warned about the misconception of facial expression and emotion detection, yet emotion determination is possible from the process and analysis of FER. FER has gained popularity in affective computing, and this anchors on the fact that the human face conveys more information than any other nonverbal communication channels [12] [13]. FER has the capacity of attributing an observable deformation in a face to a particular class of emotion [11] [12].

Emotion is also discussed in terms of dimensional space [14] [15]. This is referred to Emotion Dimension Space. A popularly known one is the arousal and valence dimensional space description of emotion. Valence describes emotion as either positive or negative, and arousal describes emotion based on the ordinal degree as high or low [16] [17]. From dimensional space description of emotion, FER inherits the task of determining emotion intensity from facial expression. Facial expression intensity estimation is an observable difference between facial expression images of the same expression or the degree of dissimilarity measures of facial expression

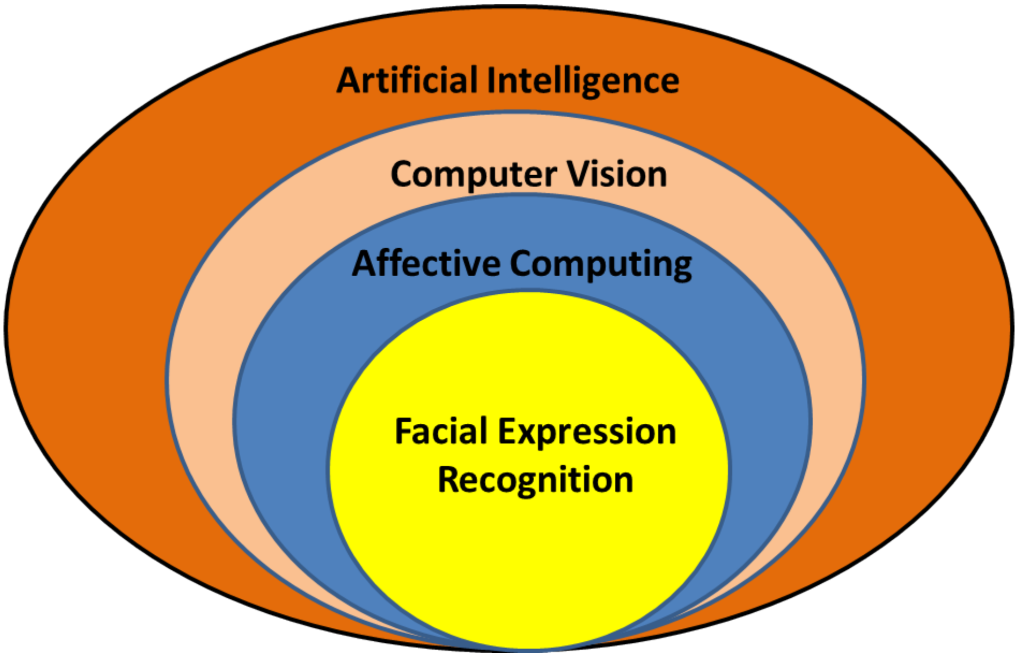


Figure 1.1: Figure describing FER as an aspect of Artificial Intelligence.

image from its reference base. Emotion detection from face and intensity estimation is very relevant to patient pain analysis [18] [19] and other complex emotion-based diseases like depression [20].

1.1.1 Facial Expression and Intensity Estimation

Expression intensity estimation is a natural human endowment. It spontaneously occurs irrespective of race, tribe and colour. Expression intensity estimation means associating a certain degree of measure with an observed emotion. Human beings are very consistent in using relative values as a metric of measurement [21]. For instance, it is very convenient for a man to express the hotness or coldness of weather with some ordinal metrics like very hot or very cold. Likewise, human beings assess expression in a face and predict an affective state of a fellow human being. Figure 1.2 illustrates how emotion is expressed differently at different times by the same subject. Accurate prediction of emotion and its intensity estimation consciously or unconsciously trigger some actions in man, which is intelligence exhibition.

Another way human beings estimate expression intensity is to infer more than one expression from a single face simultaneously. It occurs more often when the expression display is not pure [12] [21] [22]. That is, there is a mixture of basic emotions in the expression displayed. The expression intensity is estimated as "Angrily Surprise", "Fearfully Angry", "Happily Surprise", and the list continues in the combination of observed emotion. Estimating emotion in this manner reflects the ambiguous nature of facial expression images.

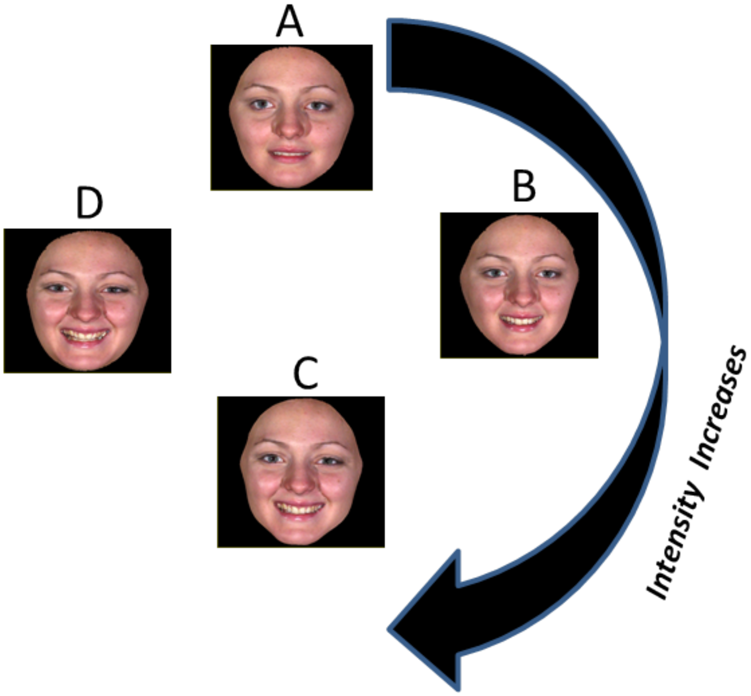


Figure 1.2: An extract from BU-3DFE dataset showing the intensity of Happy expression increasing from A to D for the same subject.

Developing an emotion recognition system also requires the consideration of intensity estimation because there will be limitations in the capacity of the FER system without considerable attention given to intensity estimation.

This research aims to automate a concurrent recognition of human emotional state with the associate intensity from facial expression images. The approach requires a good understanding of machine learning techniques, image processing, modelling, and state-of-the-art deep learning technology.

1.2 Motivation

Diverse studies are available on FER tasks, and most of the approaches regarded FER as a multiclass problem. Treating FER as a multiclass only considers assigning an emotion class to facial expression image, which has been achieved severally through handcrafted [23] [24] and machine learning models [25]. The introduction of deep learning models brings tremendous improvement to FER performance, especially when computation and optimisation options are available for insufficient data [26], which are initial sources of restricting deep learning application to FER. The multiclass approach to the FER classification task only limits the system performance because facial image contains a vast amount of information. Even the psychological research on the human affective state established that the human face rarely displays pure emotion [22] [27], which implies that expression in most times occurs as a mixture of basic emotions [28]. With this argument, categorising a facial expression into any basic emotion classes denied FER from exploring emotion-related resources from the human face.

The concept of arousal and valence in emotion has also been studied in FER [14] [15] [16]. FER recognition task based on arousal considers emotion estimation as measuring the degree or rate at which emotion is expressed in a face. Approaches employed in the literature so far are categorised in [29] into the distance-based [30], the graph-based [31], the regression-based [32], and the cluster-based [33], These methods are capable of estimating emotion using numerical values. Also, the method could perform emotion estimation either before or after recognising emotion, and some of the methods managed to perform emotion estimation without recognising the emotion [29]. Numerical estimation of emotion does not reflect human cognitive capability [28]. Man can only use ordinal metrics for measurement, which is not different from how man estimate emotion [21].

The three main aspects of FER that are researchable include the FER model, the dataset collection, and the data annotation [34]. Literature in the field focused more on achieving an optimal model [26] for emotion recognition or solving FER data related problems [27], but the aspect of data annotations has not been given much attention. Data annotation in FER is as important as both the model and the data because this determines the efficiency of FER performance and general application acceptance [34].

This research is motivated to investigate concurrent recognition of emotion and intensity estimation based on ordinal metrics. Furthermore, the investigation explores FER data annotations' correlations to resolve FER data ambiguity and annotation inconsistency.

1.3 FER Challenges

Many factors serve as limitations to the optimal performance of the FER system. The elements could be broadly grouped into class variations and person-specific identities.

In FER, classes are formed from different subjects displaying similar traits of facial deformation. This singular experimental act introduces a wide variation within the emotion classes because some information is captured with the data. FER is like other recognition systems, where intraclass variation minimization and interclass variation maximization are critical factors for system robustness. In FER, considerable intraclass variation is often observed because information about individual subjects in a class has significant differences. The difference is traceable to occlusion, light intensity variation, face morphology, gender differences, age, and facial mark (tribal or accidental) as explained below:

Occlusion An occlusion is a challenge posed due to disturbance or hindrances that obscure the characteristic feature from the expression image. This problem is not limited to natural occurrences like moustache and beard, and self-made like wearing glasses, cosmetics headscarf, or hijab.

Light Intensity Variation Illumination variation in light direction often leads to changes in light intensity and causes a cluttered background for expression images.

Face Morphology The location of a face at the time of data collection could also be a challenge in a 2D morphology; the head should be positioned in frontal view. Using 2D images reduce the computational cost, but determining appropriate facial features is extremely difficult. However, the reverse is the case for a 3D image. A side view position could affect the performance of the system. Non-frontal view and rigid head motion are challenges peculiar to spontaneous data.

Age Age categories contribute to variations in how people express emotion through the face. For example, emotions are observed in children's faces, obviously noticed in adults and mildly displayed in elders. The variations in performance were assumed to be related to infant skin texture, more fatty tissue, facial conformation, and the absence of transient furrows.

Facial Mark (Tribal or Accidental) Facial marks are some permanent deformations to the face, which may be a deliberate action in the name of beauty (tribal mark or tattoos) or accidental (facial cut from an automobile accident). Facial marks could influence FER's decision by overshadowing the facial cues needed as information by FER for prediction.

Nature of database Most Facial expression databases are collected in a controlled environment; the expression images are static, acted by either professional or non-professional actors. FER developed from a monitored environment degrades performance in a real-world where random and sequence images are available.

Similarly, the dissimilarities between classes are slight because the information collected is from the same subjects and for different emotion classes, which causes the influx of subject information between the emotion classes. The illustrated FER challenges had been the field's long-term standing challenges, generally referred to as the intraclass variation and the interclass variation. Thus, minimizing intraclass variation and maximizing interclass variation is the optimal goal in achieving a robust FER System.

Aside from the Intraclass and the interclass variations, other sources of FER challenges are the nature of FER datasets and FER data annotation inconsistencies. FER data is ambiguous, implying a possibility of having a facial expression image with attributes of two or more emotions. Then assigning a single logical label to such an expression constitutes a loss of information. Likewise, FER data annotations are conducted by experts who are human beings. The possibility of inconsistencies in the data annotation is very high due to human bias and other factors. Solving ambiguity and annotation inconsistency in FER requires a system to learn the correlation among labels and automatically annotate the data.

1.4 Aims and Objectives

This research aims at modeling a recognition framework for concurrent facial expression recognition and intensity estimation.

Objectives include:

- To investigate the current trend and techniques for FER, particularly as regards intensity estimation of emotions and data annotation inconsistencies and correlation among labels.
- To investigate the performance of a deep forest model based on forest classifiers for emotion classification.
- To investigate a deep learning-based multilabel framework for concurrent emotion recognition and intensity estimation using ordinal metrics.
- To enhance the multilabel convolutional neural network framework with transfer learning and Island loss.
- To model an efficient framework for FER label correlation and ambiguity using manifold learning and graph convolutional network.

1.5 Research Approach

The research focus is emotion recognition and intensity estimation. This task was approached first by providing a deep learning multilabel model with ordinal metrics for concurrent recognition of emotion with the corresponding intensity. This approach used other available information with the recognised emotion for the emotion estimation. The other method is the distribution learning techniques that employ graph-based label enhancement and graph convolution networks to recognise emotion and intensity based on the proportion of emotion distributions in facial images.

The multilabel Convolution Neural Network (Multilabel-CNN) model comprises of an enhanced binary relevance module and Convolution Neural Network (CNN) module. The binary relevance module ensures dependency in the data and models the ordinal metrics alongside the emotion label. Six basic emotion classes (Anger, Disgust, Fear, Happy, Sadness and Surprise) and intensity ordinal degrees (Very High, High, Normal, Low) are modelled. For effective recognition, we employ the state-of-the-art CNN module with a sigmoid activation function at the output layer of the network for the multilabel classification task. To avoid model overfitting, we employ a pre-trained optimisation technique using the VGG-16 model and mitigate the intraclass and interclass variations with island loss. Island loss minimises intraclass variation and maximises interclass variation, which optimally improves the performance of the model.

The other approach considers expression as a mixture of basic emotions. The first phase of this approach employs manifold learning to learn the correlation among facial expression emotions relying on the manifold capability to ensure that the distance of the neighbouring data along the manifold is equivalent to the corresponding distance at the edge of the graph. Before the manifold learning implementation, we compute image Euclidean distance to prevent the short circuit edge problem possibly caused by ordinary Euclidean distance. Using Isomap manifold, we compute the distance and the adjacency nodes as inputs for the graph convolution network with the softmax activation output layer. GCN is a semisupervised model which propagates the features along the graph edges and uses the distance information of the neighbouring nodes to recover the label distribution from the data with few samples from the logical labels. The resultant distribution model adequately predicts the distribution of emotion in a facial expression image and the respective proportion of their intensity.

1.6 Thesis Contributions

The main contribution of this work is to develop a FER model that simultaneously recognises emotion with the corresponding intensity from facial expression images. The following are the contributions of this thesis to the field of affective computing and computer vision:

- Chapter 2 presents a comprehensive study of the areas of application of facial expression recognition and detailed information on the discovered challenges mitigating facial expression recognition performance. Classification of existing works into machine learning approaches: single label learning, multilabel learning, and label distribution learning to indicate research trends in the field, and thorough discussion of existing methods, their performances, and their limitations are well presented.

- Chapter 3 Investigates the performance of the deep forest algorithm on FER datasets and also presents a multilabel deep learning model for concurrent emotion recognition and corresponding intensity estimation using ordinal metrics. It discusses the Multilabel CNN model enhancement with the pre-trained network that prevents the model from overfitting and the island loss function to minimise the effect of intraclass and interclass variations. It Presents a graph-based label enhancement framework called Manifold Graph Convolutional Neural Network, which employs Isomap manifold to learn correlation among data labels with the similarity distance along the manifold and recovers label distribution from logical labels using GCN.

1.7 Datasets Sources

The datasets considered for this research are benchmark datasets (Cohn Kanade extension (CK+) [2], Binghampton University-3D Facial Expression (BU-3DFE) [1]. The datasets are publicly available with release permission from their producer. We claim that the data are relevant to the challenges we considered in this work and obtained the producer's copyright permission. We also claim that we have restricted ourselves to the terms and conditions of using the data.

1.8 Organisation of work

The organisation of the thesis is as follows:

Chapter 2. The chapter covers the review and the suggestions of possible areas of FER applications and details of existing challenges in FER. It includes Categorising existing work based on machine learning problems of definitions and critical analysis. It describes various methods employed, including their performances and limitations.

Chapter 3. Presents the implementation frameworks for FER, ranging from the deep learning approach for facial expression recognition, a multilabel convolutional neural network for facial expression recognition and ordinal intensity estimation to facial expression recognition with manifold learning and graph convolution network.

Chapter 4. Contains the result presentation, analysis and discussion of each of the frameworks presented in chapter 3 including detail of the comparisons with the existing methods.

Chapter 5. In this Chapter, a conclusion with a general overview and contribution of the thesis is presented with a highlight of possible future work

Chapter 2

Facial Expression Recognition: A Review of Trends and Techniques

2.1 Facial Expression Recognition: A Review of Trends and Techniques

2.1.1 Introduction

This Section presents a holistic review study of facial expression recognition. The study captures a thorough discussion of various areas of application of facial expression recognition, systematic presentation of different factors challenging facial expression recognition, thorough discussion of facial expression trends with machine learning problem definitions, and critical analysis and presentation of existing methods for facial expression, including their performances and limitations.

Part of this work was presented in [1] and published in [2]

[1] 2019 Conference on Information Communications Technology and Society.

[2] O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," in IEEE Access, vol. 9, pp. 136944-136973, 2021, doi: 10.1109/ACCESS.2021.3113464.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 27, 2021, accepted September 13, 2021, date of publication September 16, 2021, date of current version October 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3113464

Facial Expression Recognition: A Review of Trends and Techniques

OLUFISAYO S. EKUNDAYO AND SERESTINA VIRIRI, (Senior Member, IEEE)
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4041, South Africa

Corresponding author: Serestina Viriri (viriris@ukzn.ac.za)

This work was supported by the University of KwaZulu-Natal, South Africa.

ABSTRACT Facial Expression Recognition (FER) is presently the aspect of cognitive and affective computing with the most attention and popularity, aided by its vast application areas. Several studies have been conducted on FER, and many review works are also available. The existing FER review works only give an account of FER models capable of predicting the basic expressions. None of the works considers intensity estimation of an emotion; neither do they include studies that address data annotation inconsistencies and correlation among labels in their works. This work first introduces some identified FER application areas and provides a discussion on recognised FER challenges. We proceed to provide a comprehensive FER review in three different machine learning problem definitions: Single Label Learning (SLL)- which presents FER as a multiclass problem, Multilabel Learning (MLL)- that resolves the ambiguity nature of FER, and Label Distribution Learning- that recovers the distribution of emotion in FER data annotation. We also include studies on expression intensity estimation from the face. Furthermore, popularly employed FER models are thoroughly and carefully discussed in handcrafted, conventional machine learning and deep learning models. We finally itemise some recognise unresolved issues and also suggest future research areas in the field.

INDEX TERMS Facial expression recognition, single label learning, multilabel label learning, label distribution learning.

I. INTRODUCTION

Facial Expression Recognition (FER) has gained remarkable attention in computing, which is not limited to Computer Vision (CV) and Human-Computer Interaction (HCI). The advancement in technology and the aim to achieve machine-human communication encourage many researchers to explore the field in more than two decades. FER is about detecting human affective states due to responses observed in a face through facial muscles movement due to involuntary action triggered by changes in human emotional states. From the psychological point of view, the categories of human emotional states are into six basic emotions; sad, happy, fear, surprise, anger and disgust [1]. According to the study conducted by [2], facial expression carried a larger percentage of communication information in man than any other non-verbal medium like hand gesture, body gesture, and text [3], [4]. A man without difficulty can easily interpret expression display in the face, but the automation of this task in the machine remains a challenge [4].

FER is a combination of two significant fields or disciplines (Psychology and technology). In Psychology [5], [6], facts about facial responses to emotional changes are thoroughly studied and established. Likewise, applying technology employed image processing concepts (Computer Vision) and machine learning techniques to achieve automation. FER's general architecture comprises three major phases; pre-processing, feature extraction, and classification or recognition. These phases carry out their respective tasks sequentially on a particular FER database to establish ground truth for the system to achieve its goal. Details of the FER architecture description is available in Figure 6.

FER's automation comes in two main procedures; feature extraction methods and feature classification methods. However, it is advisable to carry out some data engineering techniques before applying these methods or both accordingly. Achieving a robust system is the goal of FER. Nevertheless, FER's automation is challenged with some factors like; intensity, occlusion, facial tribal mark or accidental facial mark, face morphology, age, to mention a few. FER's emotion recognition has various applications: medicine, psychology,

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

security, clinical investigation of neuropsychiatric disorders (affective disorder or schizophrenia).

The quest for adequate recognition of man affects state led to the evolution of several approaches in developing a FER system. Existing FER review works [7]–[10] have diversely presented comprehensive studies on the traditional FER implementation methods, including the handcrafted techniques and the machine learning algorithms. Likewise, different overview studies of deep learning methods approach to FER have been presented in FER literature. The works concentrated mainly on different methods for a robust and efficient FER model. Virtually all the works provided information about the databases in the field, but studies on FER data annotations have not been given adequate consideration, which is the motivation for this work. This study considers studies that proposed methods to resolve FER data annotation inconsistency and label ambiguity. This work presents FER in three different machine learning problem definitions, which include: Single Label Learning (SLL) (Multiclass problem), Multilabel Learning (MLL): where a FER image contains one or more basic emotions. Another approach is Label Distribution Learning (LDL), which proportionally estimate all the basic emotions present in facial expression image. SLL also consider estimation of the intensity of a recognised emotion available in the expression image. No review literature in the field includes studies that consider expression intensity estimation, label discrepancies and ambiguity, and correlation among labels in their work to the best of our knowledge. The uniqueness of this work include:

- The review of existing FER application areas and suggestions of possible FER application environments to explore. This information is necessary as a quick guide or enlightenment for interested researchers in the field.
- Review of identified problems in the field that affect system performance and provides new researchers with possible challenges to consider for efficient model development.
- Review of FER literature and their classification into three groups of machine learning problem definitions, SLL: contains methods that consider FER tasks a multiclass problem. MLL: FER approach that resolves the ambiguous nature of FER data. Lastly, LDL: FER methods for label annotation inconsistency and correlation among labels.
- Provide a thorough review study on traditional FER classification models and modern deep learning models. Although literature in the field considered these separately, reviews have been presented on conventional machine learning models or deep learning models. Nevertheless, the integration of both in a single work is one of the uniqueness of this work. We purposely include them for new and interested researchers to have a general overview of what has been done in the field.

This work is organised as follows; In Section II, we discuss some FER application areas. Section III illustrates some of the challenges to be considered while developing the system to achieve a robust system with excellent performance. Section IV presents information about the available FER databases. Comprehensive FER literature studies that present FER in a different category of problem definitions are thoroughly and carefully presented in Section V. Likewise, Section VI illustrates some popularly used FER techniques in their group of handcrafted, Machine learning algorithms and state-of-the-art deep learning methods. Section VII presents a general discussion and opens up some unresolved research issues and future research areas. The last section, which is section VIII is the conclusion of this review work.

II. APPLICATION OF FACIAL EEXPRESSION RECOGNITION

There is still no limit to FER's application, and it spans through every facet in which natural interaction between man and machine is achievable. This section considers some of the areas of FER applications.

A. SOFTWARE DEVELOPMENT

The goal of every software is to meet or satisfying the requirement elicitations of end-users. Software usability is one of the means of determining the degree of satisfaction through feedback from end-users. The traditional way of measuring user satisfaction is by administering a questionnaire, but Kolakowaska *et al.* [11] believe that a questionnaire may be biased and misleading. They introduce FER as part of multimodal inputs for software usability testing and research on finding the relationship between software developers and Job's quality delivery within a particular time frame. The study's outcome shows that developers' emotions affect software productivity and quality, and they suggested incorporating an emotion detection mechanism in the HCI system.

B. EDUCATION

Education is one of the backbones of a country's economic sectors. Therefore, practical knowledge dissemination and appropriate learning are inevitable. Every institution's learning process requires thorough monitoring and proper feedback from both the learners and the instructors. The traditional methods of using surveys via questionnaire and interview have their limitations. Some factors inhibit knowledge transfer in the learning system, according to the emotional state of an individual involved [12]. These factors should be investigated regarding the assessment of learners' emotional state, evaluation of educational resources in a virtual institution and distance learning environment, and usability testing of educational tools [11]. The most appropriate means of achieving excellent results from the listed experiments would be via FER. Lisetti *et al.* [13] suggested adopting FER feedback-like mechanism into a tele-teaching assistant system in a distance learning environment and claimed that this would ensure class dynamism.

Zhou *et al.* [12] proposed an e-learning FER to capture the real-time students' emotional states for timely adjustment of teaching strategies. The recent pandemic that befalls the whole world transformed teaching and learning environments from physical contact to virtual. Most of the applications employed like zoom and the likes have the challenges of capturing students affect, which is vital information in achieving class dynamism and effective teaching.

C. MEDICINE

FER is applicable to some medical fields like; neuro-psychiatric disorder, Patients treatment feedback, patient's emotion monitoring, rehabilitation, autism and music therapy [14]. Human Facial expression has been employed in investigating neuro-psychiatric disorder as it affects emotion perception, expression and recognition in affected patients [15]–[17]. The available method used by clinicians in the field is a qualitative manual method, which is more subjective and human-intensive [16]. This challenge requires an objective process that possibly reduces human-intensive efforts and provides a qualitative result. Wang *et al.* [18] proposed the FER framework that derived probabilistic expression profiles for video data, and in turn, automatically quantified emotional expression differences between neuropsychiatric disorders patients and healthy controls. The advent of telemedicine [15], [16] in the medical field gives more justifications for FER's application. With the dynamic evolution and advancement experienced in technology development of communication devices and mobile applications such as a computer, mobile devices, video chat applications, to mention a few, could be explored using FER technology that employs facial cues to determine users' emotions in real-time.

D. SECURITY

Application of FER into identity recognition system will strengthen and improves the functionalities of the system. Biometric systems (face recognition) designs for identity authentication, and its application to security, access control, forensic and so on had been successfully achieved. Likewise, a security surveillance system saddled with the responsibility of monitoring an environment has the capability of providing detailed information on events within a specified time frame. Security surveillance System and biometric Security inclined system has the limitation of not preventing the environment from experiencing imminent attack from enemies. Adding FER to these systems will incorporate a layer of security intelligence to detect enemies' intention [19] through the emotion displays and alert the security personnel. [20] proposed improving surveillance systems by incorporating FER to make a system that would detect a person with malicious intentions from their facial expression and report to the securities before the perpetration of the intended evil. There is a need for this type of intelligent surveillance in public places like Shopping malls, Sports arenas, airports, and other places where people's gathering is encouraged.

E. MARKETING

The heartbeat of any company or business organisation is marketing, and it includes market research and advertising. The market research department could either use an interview or questionnaire, a traditional means of collecting information about users' opinions. This conventional means, according to [21], is facing out of effectiveness. Another method is to capture a user's behaviour using a sample of the product [22]. The later approach needs to carry out video analysis by experts. The method is capital and human-intensive. The cost of a behavioural approach could be minimised by employing a FER system for video analysis tasks. Yolcu *et al.* [23] developed a non-invasive deep learning-based system for monitoring customers' interest and advertisement acceptance rating. This method is more objective and reliable for adequate decision-making than the traditional way users formulate their preferences, which often mislead the research team. The advertising department could also incorporate FER into the analysis of public opinion towards various advertisement approaches. With FER, they could concentrate on the advertisement that captures more attention with positive responses.

F. ROBOTICS AND GAMES

Personal assistant robot tasks could be extended to exhibit a human-like interaction, and in most cases, they discharge their respective duties accordingly if they are embedded with a sensor that could interpret the boss's facial expression. Games or computer games should explore automatic FER thoroughly and develop game applications with characters that display affective states applicably and accordingly. It would also be of more interest if a game application could capitalise on FER for its dynamism. It should be from the user's facial expression to detect the user's feelings and trigger an action to meet the user's satisfaction.

Other areas of FER's application include image and video information retrieval, forensic investigation (Lie detector) [24], stress and depression management [25], Driver's monitoring agent in automobile [26], a fear detector at real-time in the critical mission, real-time expression recognition in mobile digital devices, temperament detection, a job interview and many more. Some of the suggested application areas have been deployed already by companies like Affestiva, EmoVu, Kairos, Nviso, Sightcorp and many more.

III. FACIAL EXPRESSION RECOGNITION CHALLENGES

FER is like many recognition systems, where intraclass variation minimisation and interclass variation maximisation are critical together with system robustness. Individual differences in a class majorly cause Intraclass variations in FER as a result of the following;

A. OCCLUSION

This is a form of the challenge posed due to disturbance or hindrances that obscure the characteristic feature from the expression image. This problem is limited to natural occurrences like moustache and beard, and self-made like wearing glasses, cosmetics headscarf, or hijab.

B. AGEING

Age categories contribute to variations in how people express emotion through the face. For example, emotional states are observed in children's faces, obviously noticed in adults and mildly displayed in elders. Cohn *et al.* [27] in their investigation on the performance of optical flow and high gradient detection algorithm on infants, the algorithm had less performance on infants compared to its performance on adults. The degradation in performance was assumed to be due to infant skin texture, more fatty tissue, facial conformation, and the absence of transient furrows. More emphasis is given by [28], [29] that different physical appearance like skin texture affects the analysis of facial expression intensity. Tian *et al.* [30] claimed that the variation in the way people express emotion could be attributed to the degree of facial plasticity, face morphology, rate of expression and frequency of intense expression.

C. POSE AND ILLUMINATION VARIATION

the location of a face at the time of data collection could also be a challenge, in a 2D morphology; the head should be positioned in frontal view, using 2D image to reduce the computational cost, but determination of appropriate facial features is extremely difficult. However, the reverse is the case for a 3D image. A side view position could affect the performance of the system. Non-frontal view and rigid head motion are challenges peculiar to spontaneous data. Illumination variation in light direction often leads to changes in light intensity and causes a cluttered background for expression images.

Aside from the intraclass problem, interclass challenges are experienced when the differences between emotion classes are less conspicuous. For instance, the same subjects are used in each of the expression classes. Interclass variation implies that the expression classes would have more similar information than unique information in the representative features.

Nature of database: Most Facial expression databases are collected in a controlled environment; the expression images are static, acted by either professional or non-professional actors. FER developed from a monitored environment are found to degrade in performance in a real-world where spontaneous, and sequence images are available.

IV. DATABASES

Facial Expression Database is a cogent and essential aspect of the FER system; like feature extraction and classifiers, facial expression database is one factor that contributes immensely to the robustness of the FER system. The early facial expression databases were posed database collected in a controlled environment [31], [32]. The choice of database for FER development depends on the type of its application. Apart from posed databases, there are also spontaneous databases captured at the real scene- a naturally expressed facial expression database. Recently, the quest to take FER beyond the laboratory to real-world applications requires facial expression databases in an unconstrained and uncontrolled environment, also termed In-the-wild databases. Figure 1 shows some selected samples of six basic image expressions from different databases. The widely employed FER databases include;

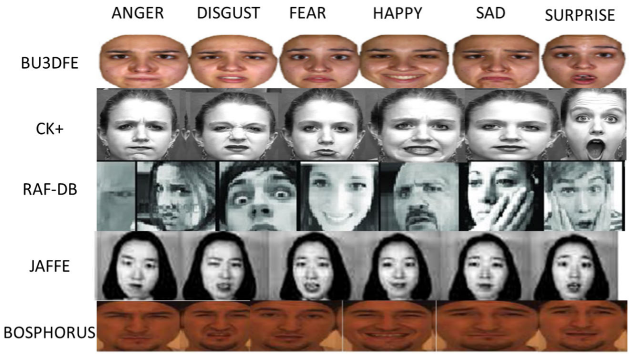


FIGURE 1. Facial expression samples of six basic emotion from different databases.

A. BOSPHORUS DATABASE

This database is one of the prevalent 3D face databases introduced by [33] and is composed of multi-expression and multi-pose facial images together with several occlusions captured in a more realistic scene. Enriched in AU and basic recognised emotional expression, adequate ground truth head pose, incorporation of different occlusion types, and employment of skilful subjects are the benefits of the Bosphorus database. Some of the compositions of this database are summarised in Table 1. The database was developed with 105 subjects altogether under different head poses, expression display and occlusion. Sixty of the subjects were men, and 45 were women, 18 wore beard/moustache, and 15 had short hair. At the point of data collection, each of 71 members of the subject had 54 face scans, and the remaining 34 Subjects had 31 face scans for each of the subjects. Despite the thoroughness exercise in capturing the AUs and the facial expression, it was not still far from the fact that they were not natural. Also, screening the AUs and the facial expressions means that not all the AUs and facial expressions will be present for all the subjects. The stated challenges are the limitations of the Bosphorus database.

B. REAL WORLD AFFECTIVE DATABASE (RAF-DB)

Raf-DB is a crowd-sourcing face data for facial expression database, categorised into basic emotions with single modal distribution and compound emotion with a bimodal distribution. According to [34] who introduced it, the database recognised it as the first of its kind, having a large scale that provided the labels of common expression perception and compound emotion in an unconstrained environment. This database's main advantages are; availability of sufficient data, no constrained or controlled environment for data capturing and group perceiving on facial expressions and data labels with the least noise. Raf-Db contains almost 30000 facial images collected with an image search API called Flickr, and the search is by using keywords relevant to each of the emotions. The extracted images were downloaded in batches using an automatic open-source downloader.

C. COHN KANADE AND COHN KANADE EXTENSION (CK AND CK+) DATABASE

Cohn *et al.* [32] released a facial expression database in 2000; the database contains 97 subjects between the ages of 18 and 30; 65% were female, and the remaining 35% were male. The subjects were chosen from multicultural people and races. There were 486 sequences collected from the subjects, and each sequence started from neutral expression and ended at the peak of the expression. The expressions' peak was fully FACS coded and emotion labelled, but the label was not validated. Luecy *et al.* (2010) itemised three challenges with CK databases; invalidation of emotion labels, Unavailable standard performance metrics for algorithm performance evaluation and lack of standard protocol for a standard database. Cohn *et al.* [35] identified the challenges with the CK database and proposed its extension, termed extended Cohn Kanade (CK+) database. In CK+, the number of subjects increased by 27%, the sequence by 22%. Also, there were slight changes in the metadata. The age group of the subject ranged between 18 and 50years. The percentage of the male and the female population is 31% and 69%, respectively. The emotion labels were revised and validated using the FACS investigator guide as a reference and confirmed by appropriate expert researchers. Leave-one-out subject cross-validation and area underneath the Receiver Operator Characteristics curve were proposed for Algorithm performance evaluation metrics.

D. JAPANESE FEMALE FACIAL EXPRESSION (JAFFE) DATABASE

Lyon *et al.* [31] introduced a database for facial expression called the JAFFE database; the database is one of the popularly used databases as a FER system benchmark. It contains ten subjects, which are all Japanese females. Each of the subjects produced 3 or 4 images for each of the six basic facial expressions. The corresponding subject images were captured while looking at the camera via a semi-reflective plastic sheet. The environment was controlled from occlusion, illumination variation, and head poses.

E. BINGHAMTON UNIVERSITY 3D FACIAL EXPRESSION (BU-3DFE)

This database was introduced at Binghamton University by [36] contains 100 subjects with 2500 facial expression models. Fifty-six of the subjects were female, and 44 were male. The age group ranges from 18 to 70 years old, with various ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. A 3D face scanner was used to capture seven expressions from each subject; in the process, four intensity levels were captured alongside each of the six basic prototypical expressions. Each expression shape model is associated with a corresponding facial texture image captured at two views (about +45° and -45°). As a result, the database consists of 2,500 two views' texture images and 2,500 geometric shape models.

F. BINGHAMTON UNIVERSITY 3D DYNAMIC FACIAL EXPRESSION (BU-4DFE)

BU-4DFE is a 3D dynamic facial expression database. The 3D facial expressions are captured at a video rate of 25 frames per second for each subject, and six model sequences showing six prototypic facial expressions. Each expression sequence contains about 100 frames. The database contains 606 3D facial expression sequences collected from 101 subjects, with approximately 60,600 frame models. Each of the 3D models of a 3D video sequence has a resolution of approximately 35,000 vertices. The texture video has a resolution of about 1040 × 1329 pixels per frame. The resulting database consists of 58 female and 43 male subjects, with various ethnic/racial ancestries, including Asian, Black, Hispanic/Latino, and White.

G. BINGHAMTON-PITTSBURGH 3D DYNAMIC SPONTANEOUS FACIAL EXPRESSION DATABASE (BP4D)

Posed and spontaneous 3D facial expressions differ along several dimensions, including complexity and timing, well-annotated 3D video of spontaneous facial behaviour is necessary. BP4D was presented by [37] as a newly developed 3D video database of spontaneous facial expressions in a different age group. The database includes forty-one subjects of 23 women and 18 men. The age ranges between 18 and 29 years; the database's cultural races are 11 Asian, 6 African-American, 4 Hispanic, and 20 Euro-American. Emotions were educed from each of the subjects using a protocol called emotion elicitation, where eight different tasks were conducted along with the interview process to deduce eight emotions.

FER databases are not limited to those discussed in this section. Information about others are briefly summarised in Table 1. [38] presented detailed information on FER databases, it's available for any interested reader.

V. FACIAL EXPRESSION RECOGNITION RESEARCH TRENDS

FER can appropriately predict individuals' emotional state from the deformation displays in the face as one of the cognitive and affective research fields. Many works have been attempted in the field to make it an achievable task. FER research has produced several models and different FER databases together with their annotations. The successes recorded so far in the literature are about FER models

TABLE 1. **The summary of some FER benchmark databases.**

Database	No of Actor	Population	Nature	Environment	Application	Available Expression
CK [32]	97	486	Sequence	Controlled (Lab)	Posed or Spontaneous	6 BExp + Neutral
CK+ [35]	123	593	Sequence	Controlled (Lab)	Posed or Spontaneous	6 Basic Emotion + Neutral
JAFFE [31]	10	213	Posed	Controlled (Lab)	Posed	6 Basic Emotion + Neutral
Bosphorus [33]	105	4888	Posed	Controlled (Lab)	Posed	6 Basic Emotion + Neutral
BU-3DFE [36]	100	2500	Posed	controlled (Lab)	Posed	6 Basic Emotion + Neutral
BU-4DFE [39]	101	606	Sequence	Controlled (Lab)	Posed and Sequence	6 Basic Emotion + Neutral
RafD [40]	67	1608	Posed	controlled (Lab)	Spontaneous and static	6 Basic Emotion + Neutral
FER2013 [41]	NA	35,887	Spontaneous	Uncontrolled (Internet)	Spontaneous and Static	6 Basic Emotion + Neutral
SFEW [42]	NA	1766	Vidoe clips	Uncontrolled (Movies)	Spontaneous	6 Basic Emotion + Neutral
AFEW [42]	NA	1809	video	Uncontrolled (Movies)	Spontaneous	6BExp + Neutral
BP4D [37]	41	NA	video	Uncontrolled	Spontaneous & Dynamic	6BExp + Neutral
Oulu-CASIA [43]	80	2880	Posed	Controlled(Lab)	sequence	6 Basic Emotion + Neutral
AffectNet [44]	NA	450,000	Wild	Uncontrolled (Internet)	Static or Sequence	6 Basic Emotion + Neutral
MMI [45]	NA	2900	Video	Controlled (Lab)	Sequence	6 Basic Emotion + Neutral
MMI [45]	25	740	possed	Controlled(Lab)	static	6 Basic Emotion
ExpW [46]	NA	91,793	Wild	Uncontrolled (Internet)	Static or Sequence	6 Basic Emotion + Neutral
RAF-DB [34]	NA	29,672	Wild	Uncontrolled (Internet)	Static or Sequence	Basic Emotion + Neutral + Compd
4DFAB [47]	180	1.8 Million	Posed	Controlled (Lab)	Static or sequence	6 Basic Emotions + Neutral
EmotioNet [48]	NA	450,000	Wild	Uncontrolled (Internet)	Static or Sequence	Basic + Compd Emotions

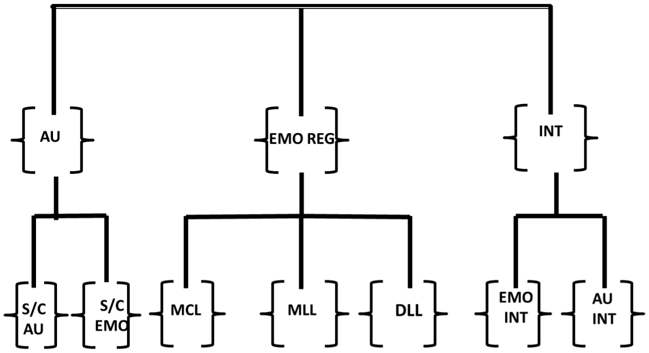


FIGURE 2. **The tree structure representing the main trends of research in FER. AU stands for Action Unit, S/C for single or compound AU, S/C EMO for single or compound emotion, EMO REG for Emotion Recognition, MCL for Multi-class Learning, MLL for Multilabel Learning, LDL for Label Distribution Learning, INT EST for intensity Estimation, AU INT for AU intensity estimation, EMO INT for emotion intensity estimation.**

that could predict the basic emotion from facial expression images. No consideration is given to other aspects of FER research that considered the intensity estimation of the emotion, Facial expression ambiguity, and the label inconsistency and correlation among labels. This section will present research diversities in FER as we categorise them based on machine learning problem definitions; SLL, SLL extension (FER and intensity estimation), MLL, and LDL. The trend in FER approaches to emotion recognition is pictorially presented in Figure 2. Table 2 presents the categories of emotion recognition research in FER with the associate limitations.

A. SINGLE LABEL LEARNING (MULTICLASS)

Early studies on the human cognitive and affective aspect of computer vision were pilots by the established work of [6], which introduced the six basic classes of emotion. Classifying an instance of face expression image into any of the six basic emotion states is identified as a multiclass task termed single label learning. Figure 5A illustrates how SLL reports only one emotion out of all the possible outcomes. Methods that attempt facial expression multiclass tasks are considerably presented in FER literature. These methods revolve around the handcrafted, conventional machine learning and the deep learning models, which we discuss in section 6. [7] is a comprehensive study of early methods on FER, [8]–[10] presented a review studies of the state-of-the-art (deep learning) methods. FER's scope as a multiclass task spreads across emotion recognition in various environments like; 1) static environment [49]–[52]. (2) Temporal and dynamic environment [53]–[55] and (3) In-the-wild [48], [56]. Several promising performances have been reported in the literature. Despite the SLL approach to FER's achievement, its simplification of assigning a single emotion to an expression instance limits its application in the real world. SLL fails to account for the inconsistency and ambiguity in FER data annotations and does not provide information about the intensity of the possible available emotions in an expression instance.

B. FACIAL EXPRESSION RECOGNITION AND INTENSITY ESTIMATION

Facial expression intensity estimation is the observable differences between facial expression images of the same

TABLE 2. Summary of diverse approaches to emotion recognition.

FER Task	Description	Limitation	Database	Metrics
SLL	A multiclass Problem that report emotion with the highest predicting value	No consideration for data annotation inconsistency, and correlation among labels	static (Bu-3DFE, JAFFE), in-the-wild, Dynamic data	Accuracy, F1, Precision, ROC
MLL	predict one or more possible emotion from the expression face	Capable of depicting the subjectivity in expression label but fails to illustrate the intensity to which each label describes the expression face.	RAF-ML, ML-JAFFE, HAPPEI, ML-BU-3DFE	RAkEL, ML-KNN, CLM, LIFT, ML-LOC
LDL	predict the possible emotion in expression image with appropriate proportion of their occurrence.	Although LDL has attempted the issue of label inconsistency and ambiguity in FER annotations to an appreciable length, yet research still open for better methods.	s-BU-3DFE, S-JAFFE	Kullback-Leibler, Euclidean distance, Sorensen, Fidelity, Intersection.

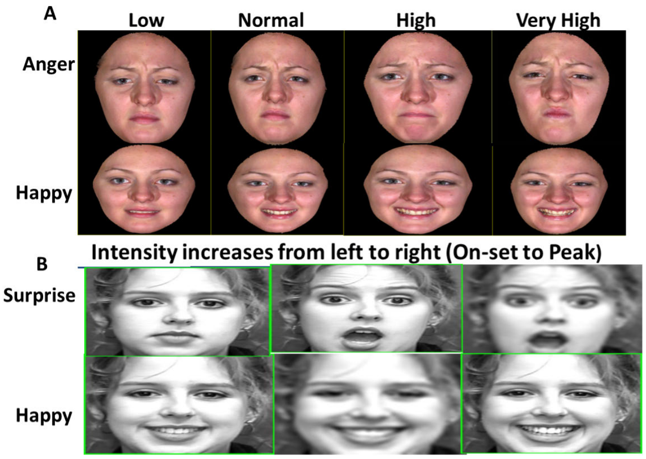


FIGURE 3. Sample A is BU-3DFE extraction (Anger, Happy) that indicates intensity displays with ordinal metrics (low, normal, high and Very_high). Sample label B is an extraction from CK+ (Surprise, Happy) showing intensity rises from ON-set to PEAK.

expression or the degree of dissimilarities of facial expression image from its reference base. One of the facial expression analysis tasks is facial expression intensity estimation; expression intensity is estimated in emotion and AUs quantifications. Figure 3 is the sample of expression intensity from static data (Figure 3A) and sequence data (Figure 3B). Some methods for FER intensity estimation have been explored in the field. Khairunmi [29] grouped these methods into; distance-based, cluster-based, regression-based, and probabilistic graphical-based.

Verma *et al.* [28] approach is a distance-based emotion intensity estimation model that uses shape transformation to capture the deformation between a template face and emotion reflected face. The deformations caused by the expansions and contractions in face regions and boundaries are quantified through elastic interpolation between the template face and expression face. The vector value generated in shape transformation is used to define a Regional Volumetric Difference (RVD) function that provides a numeric value for each of the face pixels representing the quantities of emotion displayed. Le and Xu [57] estimated facial expression intensity using isometric feature mapping. The resultant 1D manifold and facial feature trajectories are used by SVM and Cascade Neural Network (CNN) to model expression intensity. It requires that this method should conduct training for a different subject.

Observation showed that the distance-based approach quantified facial expression intensity before the recognition of the emotion. This model disagrees with how human expresses emotion.

Quan *et al.* [58] proposed a cluster-based method for expression intensity estimation. The unsupervised method employed a K-Means clustering algorithm to Haar-like features extracted from the CK+ dataset to get the K-order of the expression intensity and applied SVM classifier for the expression classification. Just like the distance-based, this approach also predicts the intensity before the expression class. Chang *et al.* [59] approach expression intensity estimation by considering the relative order information available in facial expression images. They argued that it is more appropriate and convenient to use relative order to distinguish between two expressions than considering their absolute difference. Their method employed a scattering transformation to extract discriminating and translation invariant features and used RED-SVM with Radial Basis Function (RBF) kernel for expression ranking. This method is single image-based and does not consider available temporal information.

The work of [60] is a regression-based approach, and they proposed an ensemble of naive Bayesian classifiers for expression classification and intensity estimation, respectively. They employed some naive Bayes classifiers to classify selected features weakly and generate a robust classifier from the weak classifiers' output for expression classification, and the normalised output scores are the class intensity estimation. Wu *et al.* [61] considered expression intensity estimation by quantifying energy variation of facial expression sequence. They were motivated by the possibility of quantifying energy value for each state of expression using facial landmarks. The model employed HMM to discriminate different expressions and used a linear regression algorithm to obtain intensity curves for each expression. [62] presented a regression-based model; their model utilised the ordinal information distributed in sequence image to annotate expression intensity. The proposed Ordinal Support Vector Regression model (OSVR) could generalise well in both supervised and unsupervised environments because OSVR is a combination of Support Vector Regression, which is

TABLE 3. Summary of various models for emotion and intensity recognition. NA:Not Applicable, MAE: Mean absolute error, PCC: Pearson correlation coefficient, ICC: Intraclass correlation, MAL: Mean absolute loss, HL: Hamming loss, RL: Ranking loss; AP: Average precision, CE: Coverage error.

Method	Model	DB	Performance	Limitation
Lee and Xu [57]	Optical flow tracking algorithm (Distance)	Real-time data:	88.32% (accuracy)	Need for each subject to be trained differently, not generalise, predicting intensity before emotion
Verma et al. [28]	Distance based	Primary source	NA	only few emotions are considered, method not generalise, emotion intensity before emotion recognition, computationally expensive.
Kim et al. [64]	HCORF (Prob)	CMU	89.05% (accuracy)	Intrisic topology of FER data is linearly model.
Rudovic et al. [67]	LSM-CORF (Prob)	(BU-4DFE, CK+)	(MER:19.0, 12.0), (MAL:0.36, 0.32),	Latent states are not considered in the modeling of sequences across and within the classes
Chang et al. [59]	Scartering transform + SVM (Cluster)	CK+	Mean Error: 0.313, MAE:0.318	Emotion recognition task is omitted.
Quan et al. [58]	K-Means (Cluster)	CK+	accuracy: 88.32%	Predict intensity before emotion, intensity estimation based on graphical difference is not logical
Walecki et al. [68]	VSL-CRF (Prob)	CK+, AFEW	(F1:96.7%, 28.1%), (accuracy:94.5%, 32.2%)	Result of emotion intensity is not accounted for.
Zhao et al. [62]	SVOR (Regression)	Pain DB	PCC: 0.6014, ICC: 0.5593, MAE:0.8095	Correlations between emotion classes are not modelled.
Khairuni et al. [29]	weighted vote	CK+	(Exp. acc: 82.4%), Exp. F1:69.7%, Intensity acc: 82.3%, Intensity F1: 81.8%	Emotion and emotion intensity not concurrently predicted.
Ekundayo and Viriri [69]	ML-CNN (Multi-Label)	BU-3DFE	HL:0.0628, RL:0.1561, AP:0.7637, CE:4.3140	Assume temporal information among sequence data as ordinal metrics.

responsible for intensity labels in the annotated frame and Ordinal Regression, a baseline for temporal order for frame sequence and not the label intensity values.

Probabilistic graphical-based model for emotion and intensity estimation have been thoroughly reported in [63]–[66]. [63], [65] used HMM and CRF to successfully recognise the emotion or the intensity of the target expression. [64] identified the limitation of the existing models and enhanced the discriminative ability of CRF. They proposed a Hidden Conditional Ordinal Random Field (HCORF) model to simultaneously capture multiple emotions and their respective intensities. Despite this improvement, HCORF is limited to the variations in facial expression and their respective intensities, which a simple linear model could not adequately express. Rudovic *et al.* [67] enhanced the capability of HCORF; they used ordinal manifold, a low dimensional manifold to model facial affective data topology and incorporated it into the HCORF model. The ordinal manifold preserves facial expression discriminative information and the ordinal relationship of the corresponding intensity. Walecki *et al.* [68] complimented laplacian shared parameter Multi-output CRF and HCORF and proposed Variable-state Conditional Random field method, which considered both nominal and ordinal latent state in the model of expression sequence both within and across the expression classes. They reported that the proposed method outperformed HCORF and LSM-CRF but failed to state the intensity estimation result categorically. Khairuni [29] introduced a method that employed weight voting and Hidden Markov Model for expression recognition and intensity estimation. HMM is saddled with detecting the input frame's emotion in the method, and change-point detection captured the temporal segment. The result showed that the proposed method performed better than any existing probabilistic graphical methods in accuracy and computation time. Our approach to FER and intensity estimation is presented in [69]. We considered FER and intensity estimation a multilabel task with the motivation that an instance of a facial expression image contains information about emotion displays and the corresponding intensity. We proposed ML-CNN (Multilabel Convolution Neural Network) that uses CNN as a binary classifier for an enhanced binary relevance model. We optimised the model with a VGG-16 pre-trained network and employed island loss to minimise intraclass and interclass variations. Our model concurrently predicts emotion and its intensity using ordinal information available in the data. The predictions of our model are presented in Figure 4. The experiments conducted on BU-3DFE and CK+ datasets produced an optimal result.

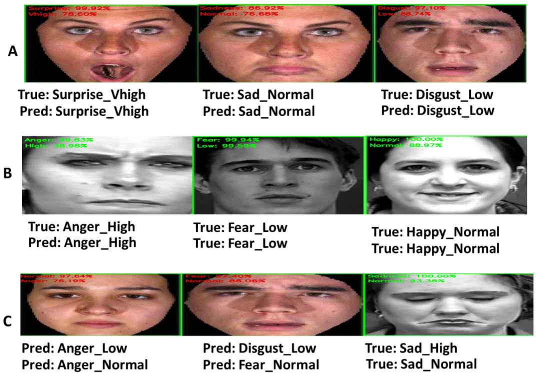


FIGURE 4. ML-CNN concurrent predictions of emotion with the associated intensity of BU-3DFE and CK+ test datasets. A and B are samples of correct predictions, and C is the samples where one of either the emotion or the intensity is incorrectly predicted.

The summary of the models for emotion and intensity estimation and their corresponding evaluation are presented in Table 3.

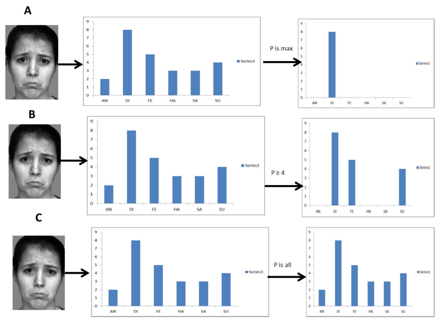


FIGURE 5. A is the description of FER Multi-class learning, where only the class with the highest prediction value becomes the identified expression. B is a FER multi-label learning scenario where more than a class with prediction value equal to or greater than a certain threshold. In C (FER distribution learning), all the expression classes are identified along with their respective prediction values.

C. MULTILABEL LEARNING

Ekman *et al.* [1], and Plutchik *et al.* [70], [71] reported that facial expression is more of a mixture of basic emotions and that a single basic expression is only displayed on a rare occasion. The argument defines the FER task as a Multilabel (ML) problem. Figure 5B shows multilabel prediction's possible output. An instance of expression image could contain one or more basic emotion information in facial expression multilabel tasks. There are few FER literature with a multilabel approach; this resulted from the few available multilabel datasets. The datasets list include; JAFFE [31] BU-3DFE [36] HAPPEI [72], EmotioNet [48] and the most recent RAF-ML [34]. One of the multilabel methods applied to FER is Group Laso Regularised Maximum Margin classifier (GLMM) proposed by [73], GLMM considered the fact that the AU at different affective states is triggered in the same region of the face. GLMM used the feature extracted for different expressions at the same region to classify them into a zero or non-zero, making it possible for a group to contain different expressions. The global solution of the model was achieved by a function called Maximum Margin Hinge loss. GLMM was later enhanced to Adaptive Group Lasso Regression [74] to assign a continuous value to the distribution of expression present in a non-zero group. GLMM shows its superior performance compares with some existing ML methods from the experiment conducted on s-JAFFE. The work of [34] is also a multilabel approach to FER, Li and Deng [34] introduced a multilabel deep learning model termed Deep Bi-Manifold CNN (DBM-CMM). The model preserves the local affinity of deep emotion features and the manifold structure of emotion labels, while learning the discriminating feature of multilabel expression. The deep network training is jointly supervised by softmax cross-entropy loss with the bi-manifold loss for feature discriminating enhancement. This model learned emotion distribution properly from RAF-ML data and generalised well with existing multilabel data through the incorporated adaptive mechanism.

D. LABEL DISTRIBUTION LEARNING

The extension of the multilabel approach is the Label Distribution Learning (LDL). The main reason that triggers the introduction of the LDL approach to FER is the inconsistencies in FER datasets annotations, which might be due to human annotators' subjectivity, and the subtlety and ambiguous nature of FER data [75]. These challenges adequately justify the need for LDL because LDL could assign multiple labels in different proportions to an expression image. One of the LDL application studies to FER is Emotion Distribution Learning (EDL) [71]. Ying *et al.* [71] resolve the challenge of emotion intensity information loss in the SLL and MLL approach and propose the EDL method to eliminate the threshold constraint. The EDL method describes emotion intensity as a probability distribution of basic emotions present in facial expression, and finally assigns each emotion to the computed degree of intensity. EDL outperform some existing LDL methods and MLL methods when evaluated on s-JAFFE and s-BU-3DFE datasets. In the same manner, [76] proposed two LDL models, which are LDLogitBoost that employs weighted regression tree as the base learner and AOSO-LDLogitBoost that uses vector as base learner. These algorithms are Logistic Boosting Regression (LBR) Based formed from additive weighted function regression. Both LDLogitBoost and AOSO-LDLogitBoost show a promising performance when evaluated on s-BU-3DFE. This method only considers data with distribution scores.

Similarly, [77] proposed an EDL method based on surface Electromyography (sEMG) that uses PCA as feature selection and Jeffery's divergence to find similarities between basic emotions. The sEMG based distribution learning system gains from the robustness of EMG features to head pose variation, the possible influence of external factors, and their unbias information. Nevertheless, EDL is only applicable to datasets with emotion distribution scores.

Most FER databases do not come with distribution scores; applying LDL to these datasets requires methods to recover or relabel the data with distribution scores. The few techniques that consider this challenge include; label enhancement based on fuzzy clustering algorithms [78], which employs C-means clustering to cluster feature vectors and iteratively minimise the objective function to achieve label distribution from logical labels. Another group of Label enhancement is graph-based label enhancement, which includes enhancement algorithm based on label propagation [79] and manifold learning algorithms [80]. The motive behind manifold learning-based label enhancement could achieve label distribution by reconstructing every data point from its neighbour through graphical representation of the topological feature space. In comparison, label propagation-based label enhancement depends solely on iterative propagation techniques to generate label distribution from the logical label. These methods create the distribution labels, but they fail to consider the correlation among the labels. [81] approach distribution label recovery with Graph Laplacian Label Enhancement (GLLE)

TABLE 4. Analysis of MLL, LDL and label enhancement models.

Author	Method	Data	Performance Evaluation	Contribution	Limitation
Zhao et al. [74]	GLMM	ML-JAFFE	Average Precision: 0.9143, Coverage error:2.9381, hamming loss:0.2035, One error:0.1071, ranking loss:0.1466	Model the relationship among FER labels	Model not capture the intensity estimation of the available emotions in FER data.
Ying et al. [71]	EDL + JD	(sJAFFE, sBU-3DFE):	(Kullback leibler:0.0346,0.0402), (Euclidean:0.0957,0.1005), (intersection:0.8998,0.8939), (fidelity:0.9914,0.9898)	present information about emotion intensity in an expression instance	limited to label distribution data.
Xing et al. [76]	LDLogitBoost	s-BU-3DFE	Kullback Leibler:0.0491, Euclidean:0.1263, Fidelity:0.9886, intersection:0.8800	more general entropy model for modeling information distribution in facial images	Not generalised to in-the wild and logical label data, the performance may degrade with large volume data
Xing et al. [76]	AOSO-LDLogitBoost	s-BU3DFE	Kullback Leibler:0.0515, Euclidean:0.1297, Fidelity:0.9874, inersection:0.8764	more general entropy model for modeling information distribution in facial images	Not generalised to in-the-wild and logical label data, the performance may degrade with large volume data
Li and Deng [34]	DBM-CNN	RAF-ML	CLM-Hamming: 0.217, RAkEL-Hamming:0.177, ML-KNN-Hamming:0.168, ML-LOC:0.173, LIFT-Hamming:0.167	Preserves both local affinity and manifold structure of emotion label. Introduction of Adaptation mechanism for data generalisation	computational complexity and resource consumption
Xu et al. [81]	Label enhancement with GLLE (manifold learning)	Bu-3DFE	cheb:1.00, clark:1.13, canb:1.13, cosine:1.00, Interception: 1.07	Label enhancement with consideration given to correlation among labels. Could be applied to data with no distribution label	Not advisable to use on large data size or in-the-wild data. It is Computationally expensive due to implementation of KNN search.
Jia et al. [82]	EDL-LRL + ADMM optimiser	(s-JAFFE, S-BU-3DFE	(cheb: 0.0806,0.0951), (clark:0.3008,0.3556), (cand: 0.6134, 0.7463), (Kullback Leibler:0.0361,0.0694), (cosine: 0.9660,0.9626), (intersection: 0.8970:0.8686)	Preserve correlation among data label locally.	Not generalised to in-the-wild data and data with logical label
Abeere et al. [84]	EDL-LBCNN (CNN + LBC features) KL loss	s-JAFFE	Kullback Leibler:0.0168, CS:0.9842	system performance increases via hybrid convolutional features.	Not generalised to in-the-wild data and data with logical label
Zhang et al. [83]	Cosine similarity + Deep CNN	Oulu-CASIA NIR FER	(Accuracy 81.97% in weak light), (Accuracy: 82.67% in the Dark), (Accuracy in strong light: 84.40%)	the model is immune to illumination variations	not applicable to data with logical label and not generalises to in-the-wild data.
Chen et al. [75]	Auxiliary label (manifold learning) + Deep CNN	posed data(CK+, Oulu-CASIA, CFEE, MMI), wild data (AFFNET, RAF, SFEW)	(Avg. Accuracy: 76.25), (Avg. Accuracy:66.64)	resolve label inconsistency using label enhancement with correlation among labels. Applicable to data without distribution label, not affected by data volume, minimises searching with approximate kNN. Generalises to in-the-wild data and logical data	Time complexity and resource consumption due to auxiliary label space construction.

method. This method successfully generates distribution labels by leveraging topological information of the feature space and adequate consideration of the correlation among labels with appropriate optimisation. GLLE outperforms almost 11 different ML methods, based on the experiments' report on the BU-3DFE dataset as one of the datasets considered. GLLE application to a large dataset and FER data in the wild fails because of its profound assumption of topological space and K- Nearest Neighbour (KNN) search implementation.

Recently, there has been a considerable increase in the quantity and number of FER databases, encouraging the state-of-the-art method, deep learning, for emotion recognition. Using deep networks for distribution learning in FER is evident in [75], [82]–[84]. Jia *et al.* [82] in their quest to preserve the correlation among FER data label locally, they proposed EDL-LRL (Emotion Distribution Label-Low Ranking label correlation Locally), which forms a low-rank structure that alleviates the complexity in emotion correlation, with an assumption that low-rank structure represents the label space. The experiment conducted on label distribution datasets (s-JAFFE and s-BU3DFE) shows the proposed model's prominence. The model considers the correlation among the label locally on data with a distribution label. A generalisation of the method to in-the-wild data and data with a logical label is a challenge. [75] generate an auxiliary label space from two different tasks with intimate correlation with facial expression recognition. The auxiliary tasks employed are facial landmark detection and action unit recognition, which depend on facial structure and movement. This method's motivation is the possibility of two expression images in the auxiliary label space having close expression distribution and consistency in their annotations. This method minimises the problem encountered in GLLE for label enhancement by using approximate KNN for building the approximate KNN (akNN) graphs that generate the auxiliary

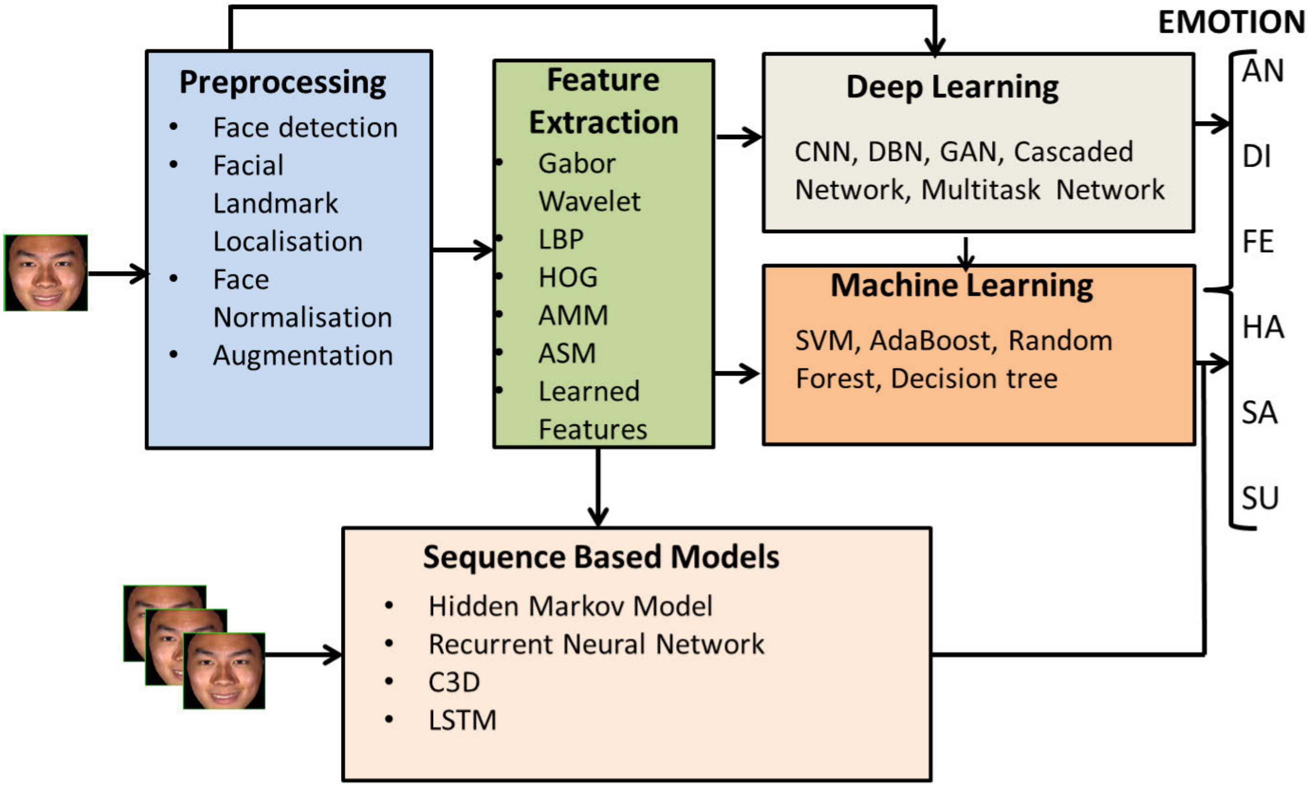


FIGURE 6. Facial expression recognition framework.

labels. Deep CNN was used as the backbone of the proposed system. An experiment conducted on laboratory-controlled data (CK+, Oulu-CASIA, CFEE, MMI) and in-the-wild (AFFNET, RAF, SFEW) proved the system's efficiency over existing methods with an assurance of label consistency and removal of label ambiguity. Zhang *et al.* [83] proposed a Correlated Emotion Label Distribution Learning (CELDL) model for Infrared facial expression recognition. The model initially computes the correlation between expression images using cosine similarities and finally learns the basic emotion in infrared expression with deep CNN. [84] proposed a feature hybrid based model called EDL-LBCNN, which hybridised Local Binary Convolution (LBC) features and Convolution Neural Network (CNN) features train with Kullback-Leibler loss and optimise with ADMM (Alternating Direction Method of Multipliers). The outcome of the experiment on the s-JAFFE dataset shows its promising performance. Figure 5C represents the LDL approach to FER, and Table 4 provides information about the MLL and LDL FER models.

VI. FER ARCHITECTURE

Although FER architecture contains two significant phases, the feature extraction phase and the classification or recognition phase, in most cases, the preprocessing stage is a crucial phase that should not be left out. Automatic FER architecture most time begins with the preprocessing phase. Figure 6 contains major preprocessing algorithms employ in FER. The next phase is feature extraction, where the discriminating features are extracted using feature extraction techniques. The last phase is the classification phase, where each expression image belongs to one of the six basic classes of emotion.

A. PRE-PROCESSING PHASE

Facial feature preprocessing is a vital phase in FER. It assists in preserving relevant features by limiting the infiltration of redundant information during data extraction. It has been observed that data preprocessing has a significant influence on the performance of both conventional machine learning methods and deep learning models. Several algorithms have been proposed in FER, and the list is not limited to face localisation, facial landmark localisation, face normalisation, and data augmentation. We shall elaborate briefly on each of the listed preprocessing methods in the subsections below.

1) FACE LOCALIZATION

Face localisation algorithms help detect the region and the size of a human face in an image or frame of images. It removes the possible background information that may influence the prediction of FER. One of the most popularly used methods for face detection is the algorithm proposed by [85]. They employed Haar-like features and used the AdaBoost classifier to learn a strong classifier from weak

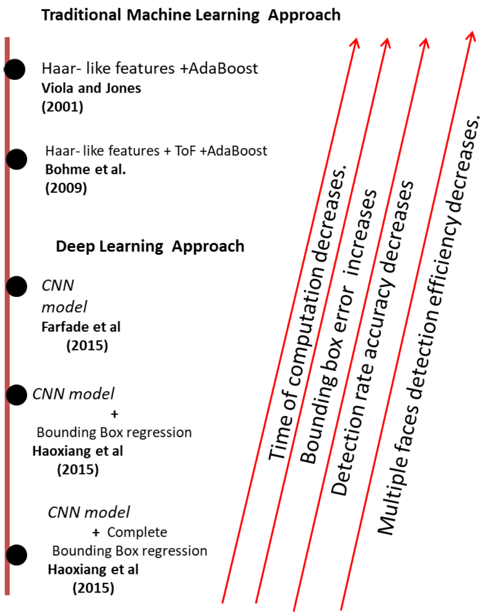

Traditional Machine Learning Approach
Haar- like features +AdaBoost
Viola and Jones
(2001)
Haar- like features + ToF +AdaBoost
Bohme et al.
(2009)
Deep Learning Approach
CNN
model
Farfade et al
(2015)
CNN model
+
Bounding Box regression
Haoxiang et al
(2015)
CNN model
+ Complete
Bounding Box regression
Haoxiang et al
(2015)
Time of computation decreases.
Bounding box error increases
Detection rate accuracy decreases
Multiple faces detection efficiency decreases.

FIGURE 7. Analysis of face detection algorithms.

cascaded classifiers. The algorithm was optimised for speed with integral images.

The study conducted by [86] showed that the model proposed by [85] outperformed LBP-AdaBoost, GF-SVM and GF-NN methods in both speeds of computation and detection accuracy. Despite the excellent detection rate achieved by this algorithm, the training cost is considered expensive. Other identified shortcomings of the Viola and Jones method include non-robustness to partial occlusion and limitation to angular face position. Bohme *et al.* [87] enhanced the Viola and Jones algorithm with range and intensity data from the Time of Flight (ToF) camera. The report showed that the improved method gained a better detection rate at a reduced training time. [88] proposed a CNN model to minimise face detection algorithms' limitation to face angular position, which efficiently detects multiple faces in diverse poses, illumination, and occlusions.

Nevertheless, the method fails to implement bounding box regression. Haoxiang *et al.* [89] worked on the deficiency of [88] model and introduced a cascaded CNN based model that employed bounding box regression. This method failed to fully utilise bounding box regression because it did not evaluate the bounding box for possible reuse. Luo *et al.* [90] fully explore bounding box regression in their CNN model for face detection to determine if the bounding box is fit for a face. They iteratively applied bounding box regression until achieving the appropriate fit and face localisation begins the preprocessing stage of FER architecture. Figure 7 presents an overview of the face detection algorithm discussed.

2) FACIAL LANDMARKS LOCALISATION

Facial landmark localisation (facial alignment) has gained remarkable popularity in Computer Vision and Biometrics. Facial landmarking requires a face detection algorithm before its implementation. The available facial components coordinates (eye-brows, mouth corners, nose ridge, eyes, and lips) in facial landmarks could improve a FER system due to their tendency to minimise in-plane rotation variation. Before the dominance of deep learning methods, most literature employs facial alignment for feature extraction enhancement. Happy *et al.* [91] reconstruct facial patches position in the face using a facial landmark detection model and edge detection algorithm. Happy *et al.* [91] used the method for the extraction of distinctive active patches for expression recognition. [92] before extracting feature patches with HOG, they first located 68 facial landmarks using ensembles of regression trees; some of the points generated formed the patches extracted. A comprehensive study conducted on facial landmark localisation is available in [93], for any interested reader. In recent years, deep learning-based models have been frequently adopted for facial landmark detection. The models have proved their superiority over other models in every facial landmarking detection competition [54], [73], [94]. The authors implemented a combination of cascading CNN modules with specific modifications to the network proposed by [95] that employed three different cascading modules to predict five landmarks. Bodini *et al.* [96] contain more information on deep learning-based methods for face landmark localisation. This paper will consider some literature that applied facial landmark detection at the preprocessing phase in a FER model.

Zhu *et al.* [97] introduced a CovNet model that incorporates the face landmark detection method proposed by [98], which produces 68 fiducial points in the face. The landmark detection aid in the creation of images with eye-brows and mouth locations. The model's performance ascertained the claim that facial landmarks position and shape representation learning could improve expression recognition from images. [99] considered AAM to generate a transformed face region of bidirectional warping facial landmarks for face registration, and that precedes the CNN and Conditional Random Field (CRF) model in solving FER task in a Spatio-temporal environment. [100] employ the supervised descent method to track 49 facial landmarks on facial expression frames in the wild, which could be used by both handcrafted methods and the DNN models for facial expression classification. Many deep learning models supported the prospect of facial alignment in FER, especially in the Spatio-temporal environment.

3) NORMALIZATION

Face normalisation algorithms tend to compliment the effort of face localisation and face alignment. It is expedient to use face normalisation algorithms after facial alignment so that problems that are feature independent (rotation, brightness, background and occlusion) could be minimised. The types of available face normalisation include; geometric, lighting, head rotation (Head Pose), face expression and occlusion. The application of the list depends on the challenges involved. The lighting and the pose normalisation are necessary for FER in an uncontrolled environment. Variation in the illumination of faces is a significant problem in FER because there is a high tendency for images of a particular subject to differing in brightness and contrast. The lighting normalisation approach minimises intraclass variation that arises from the lighting condition. Li *et al.* [101] use homomorphic filtering normalisation, a photometric normalisation algorithm and histogram equalisation for face preprocessing and claimed that the combination of the two techniques produced effective performance. Shin *et al.* [102] conduct experiments on four different lighting normalisation algorithms, histogram equalisation, isotropic diffusion-based normalisation, DCT-based normalisation, and Difference of Gaussian (DoG). Results showed that the deep network that employs Histogram equalisation at the preprocessing phase has outstanding performance compared to the same network that implements other methods. Bargal *et al.* [103], and Pitaloka *et al.* [104] are among other works that employ histogram equalisation at the preprocessing stage with promising performance. Histogram equalisation normalisation works best when the face foreground and the background are nearly uniform in brightness. Otherwise, local contrast emphasis is possible to occur [105]. Kuo *et al.* [105] proposed combining histogram equalisation and linear mapping to solve the problem of local contrast emphasis. Another hindrance to FER optimal performance in an uncontrolled environment is head pose variation. Pose normalisation has been used severally in the literature to neutralise the pose variation effect.

Most approaches to pose variation correction involve 2D and 3D model fitting that incorporates facial alignment [106], [107]. The motive behind the 2D model fitting for pose variation is that desire pose could be achieved by warping the face with 2D geometrical transformation, the methods that use 2D model fitting techniques capitalised on the working of the facial landmarking method and the warping algorithm.

Sagonas *et al.* [108] generated a frontal face image by applying a Robust Statistical based method. [109] enhanced AMM-based approach for facial landmarks, which, in turn, enhances the fitting process initialisation. The use of the Discriminating Appearance Model (DAM) for pose normalisation is considered in [110]. [111] addressed pose normalisation using Gaussian Process Regression (GPR) and affine transformation. The 3D model fitting achieves normalisation in three procedures; (i) fitting a 3D model on a located facial landmark. (ii) Mapping of face texture to the landmarked 3D model. (iii) Generation of the desired facial pose image from the 3D model texture. 3D model fitting for pose normalisation has been explored diversely in the literature. [112] used a five landmark-based 3D model and quotient image symmetry to develop a lighting aware pose normalisation. [107] introduced a homographic-based pose normalisation technique from the dense grid-based 3D landmark. [113] proposed a method that employed 3D Morphable Model (3DMM) and an interpolation method for frontal view reconstruction. Likewise, [114] synthesised frontal face using 3D Generic Elastic Model (3DDEM) with texture mapping. [115] generate a frontal face from five facial landmarks 3D mesh in a single reference. Deep learning models also explore 2D and 3D model fitting for pose normalisation. The deep learning model was able to synthesis frontal faces from the training of several multi-posed data. [115] used the deep learning method to achieve pose and illumination normalisation, and they trained a deep neural network with face images generated from 3DGEM. [116] introduced the Face Frontalization Generative Adversarial model (FF-GAM) using 3DMM. Model fitting approach for pose normalisation is expensive in terms of time and computational resources. Instead of fixing pose variation with a model fitting method, Obaydy [117] presented a technique that fully utilised facial landmarking and thin-plane spline warping technique for face normalisation, and they were able to efficiently produce a frontal face image from pose variation image in a video. Table 5 contains some normalisation models for FER and their target challenges.

TABLE 5. Presentation of normalisation models application to FER and the target challenges.

Model	Author	Year	Robustness
GPR +Affine transformation	Yang et al. [111]	2010	Pose variation
Histogram Equalization + Hormomorphic filtering	Li et al. [101]	2015	Light variation
DAM (Discrimination Appearance Model)	Gao et al. [110]	2015	Pose variation
Histogramm Equalization	Bargal et al [103]	2016	Light variation
Facial landmark& enhanced AAM	Haghigat et al. [109]	2016	Pose variation
Homomorphic Normalization + 3D landmark	Yao et al. [107]	2016	Pose variation
3DMM + interpolation frontal view	Ferari et al. [113]	2016	Pose variation
3D Mesh facial landmark	Wu et al. [115]	2016	Pose variation
Histogramm Equalization	Pitaloka et al. [104]	2016	Light variation
Satistical based model	Sagonas et al. [108]	2017	Pose variation
FF-GAN	Tran et al. [116]	2017	Pose variation
landmark + 3D + quotient Image symbol	Deng et al. [112]	2017	Pose variation
3D Generic Elastic model	Mendoza et al. [114]	2018	Pose variation
Histogram Equalization + linear mapping	Kuo et al. [105]	2019	Local contrast emphasis
Facial Landmarking + Warpping	Obaydy et al. [117]	2019	Pose variation

4) AUGMENTATION

Data augmentation is a policy adopted in computer vision to improvise for data limitation, a long-time challenge in the field. Data augmentation alleviates data challenges in deep learning through computational manipulations like flipping, cropping, scaling, rotation, and many more. Data augmentation has a significant contribution to machine learning models' performance, especially the deep learning models. Implementation of data augmentation could be done by

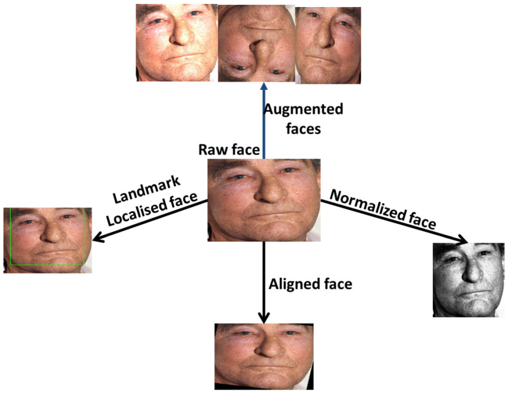


FIGURE 8. The display of some preprocessing algorithms output (augmentation, face localisation, landmark localisation and normalisation) on the raw image.

offline approach or by online approach. The offline method is employed when training data is of few hundreds, while the online approach augments data on the fly. Data augmentation has been widely explored in works of literature [118] and notable in FER [119] where there is a need for large data size. Some works consider the automatic augmentation policy learning approach because of possible biases introduced into the dataset due to the wrong augmentation policy.

Among the existing augmentation policies learning approaches [120]–[124] the work of [125] is the state-of-the-art. [125] introduce AutoAugment using reinforcement learning as a searching technique for augmentation policy with an associated probability. The result guides the system to decide the required policy that is appropriate for the dataset. Cubuk method achieved a significant efficiency, but the reinforcement searching algorithm makes the method computationally expensive. The augmentation policies learning method proposed by [126] is called Population-Based Augmentation (PBA) schedule. This approach generates an augmentation schedule from the Population-Based Training (PBT)- algorithm introduced by [127]. The method is both time and computationally cost-effective compare to the state-of-the-art. In computer vision, a robust augmentation policies learning method is still open research. Figure 8 presents the transformation that occurs after the application of any of the preprocessing algorithms discussed.

B. FEATURE EXTRACTION

As mentioned earlier, the human face is an embodiment of information. A facial image is represented with vast, complex numeric data void of human understanding. The subject information in the image data is termed feature, and extracting useful features from image data correctly to preserve accuracy is called feature extraction. Feature extraction is usually downsized or causes a dimensional reduction in a dataset because it removes redundant attributes from the data to prevent computational complexity, overfitting, and non-generality of the feature model. Every feature extraction technique's main goal is to achieve a feature representation with minimum intraclass variation and maximum interclass variation of high discriminating features. For a FER task, the popular feature extraction techniques include the appearance-based method, geometric-based method, learning features-based method and the hybrid-based method. Each of these methods contains different algorithms for feature descriptors. Both the appearance-based models and geometric based models are classified as handcrafted feature models.

1) HANDCRAFTED FEATURE MODELS

Appearance-based models could describe facial expression features as either a global feature or a local feature. This method extricates changes in the facial image by convolving either the whole image or some region of interest in the image with an image filter or filter bank [128]. Global feature descriptor algorithms translate image features into a single multidimensional feature vector of either colour, shape or texture. While the local feature descriptor algorithm is more concerned about interest points (key points), the number of interest points N forms the N-dimensional feature vectors. The following are the famous appearance-based feature extraction algorithms for FER.

a: GABOR WAVELET

This descriptor was named after the man called Denis Gabor by 1946 [129]. It is a local descriptor. Gabor's image analysis finds a region in the image with a specific frequency in a particular direction; the frequency and orientation description made Gabor appropriate for image texture representation and discrimination. A Gabor filter is a function obtained from amplitude modulation of a sinusoid with Gaussian function in a spatial domain and captures the relevant frequency spectrum. The strength of the Gabor wavelet transform algorithm for feature extraction is its adequate directional selectivity, spatial and frequency maximisation of information, and sensitivity to a slight shift in direction. Equation (1) is the formal definition of the Gabor filter. Assuming the following parameters: (x,y) to be the pixel position in the spatial domain, α to be the wavelength in pixel, θ to be the orientation of the Gabor filter and Sx, Sy to be the standard deviation along the x and y direction; then:

$$G(X,Y) = \frac{1}{2\pi S_x S_y} \exp^{\left[-\frac{1}{2}(\frac{x'^2}{S_x^2})+(\frac{y'^2}{S_y^2})\right]} \exp^{\left[j\frac{2\pi X'}{\lambda}\right]} \quad (1)$$

where X' = xcosθ + ysinθ and Y' = − sinθ +ycosθ

Lajevardi [130] argued that the whole Gabor feature extraction method both consume time and yield highly dimensional feature vectors. They considered proposing an average Gabor filter feature method that reduced the feature samples for each facial image from 491520 samples to 12288 samples before downsampling and applying PCA for dimensionality reduction. They achieved this by decreasing 40 feature images of size 128 × 96 pixels each to an average feature image of size 128 × 96 pixels. They used 64 samplings and PCA for dimensionality reduction and applied the K-means classifier on the finally extracted fea-

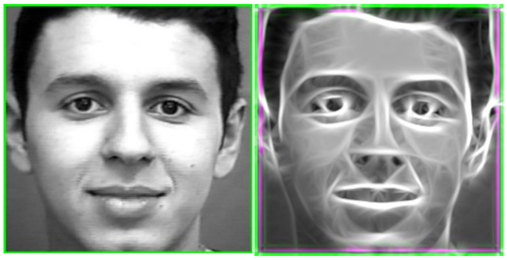


FIGURE 9. The raw image is placed at the left, follows by the Garbo filter version.

ture. The experiment's result on the JAFFE database showed that the method almost has the same output as the fully Gabor filter method and a gain of minimum time and space consumption. The work of [131] was similar. Still, instead of averaging as in [130], they proposed superimposition of eight images generated from each facial expression image when eight orientation Gabor filters were applied to obtain a single Gabor filter transformation image. Sisodia *et al.* [132] in an approach to minimise computation complexity and dimensionality reduction, selected the best representative number of significant Gabor features to represent each image's expression. Recently, Verma *et al.* [133] follow after the work of [134] however, the significant difference is that [133] employed Gabor filters for feature extraction. They used a Gabor filter bank of five frequencies and eight orientations to convolve each expression image, producing 40 Gabor magnitude images as the required Gabor feature. Harit *et al.* [135] used a Gabor filter to extract features from a normalised face for fiducial points detection. He initially created a Gabor filter bank of six orientations and three spatial frequencies, and later convolved each point in the image with the created filter bank. They reported that 1224 features extracted from each image were generated from 18 magnitudes from 68 fiducial points in the expression image. The classifier used is ANN; the experiment was conducted on two different datasets; JAFFE and Yale. The results showed that the method performed better on the JAFFE database having 81% accuracy, than Yale with 57% accuracy. The Gabor filter transformation of a happy expression image is available in Figure 9.

b: LOCAL BINARY PATTERN (LBP)

Ojala *et al.* [136] proposed LBP as an image texture algorithm suitable for texture analysis. The motive behind the LBP descriptor is that image texture can be represented by the local spatial, with a high tendency to benefit from the grayscale contrast [136]. LBP operates on each of 3 × 3 pixels of grayscale values of an image and thresholding every neighbour pixel P(0, ..., 7) with the centre pixel R(1) to generate a binary sequence using a binary threshold function S(x) and then compute the decimal equivalent for the centre pixel with (2).

$$LBP_{R,P} = \sum_{p=0}^{p=1} S(x)2^{p} \qquad (2)$$

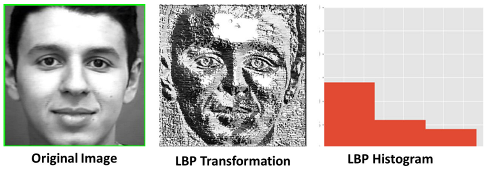


FIGURE 10. The raw image on the left, the LBP version of the image is at the centre, and the histogram of LBP image is on the right side.

Both the LBP equivalent and LBP histogram of a happy image are presented in Figure 10.

The histogram of the LBP encoded region is computed and use as a texture descriptor of that region. The strength of LBP for texture analysis lies in its tolerance for monotonic illumination changes, pose variation and computational simplicity. LBP and its variants have been widely explored in FER. LBP was used as representing feature for FER by [137]–[140]. However, the LBP feature descriptor results in poor performance in the presence of noisy data. This is because it concentrates only on the signs of the difference between the gray values and considers the magnitude relevant texture information as irrelevant. [4], [128] enhanced LBP with a feature selection algorithm. LBP pattern was extracted by dividing a facial image into regions, then the histogram of each region was calculated and later concatenated to form a single face image vector. The feature selection algorithm is further applied, which derived LBP images for all available images and groups all the images into their respective expression classes. Pixel's variance is then computed for each image in the expression class. A threshold called average variance was set to capture high variance code and low variance code. The binary image was formed from the high and low variance code matrix union and became the reference feature selection for LBPs. The experiment conducted on BU-3DFE showed better performance as reported in [4].

Ahmed *et al.* [128] understood the challenges with the original LBP and proposed a Compound Local Binary Pattern (CLBP)- a variant of LBP that uses 2P bits instead of a single P bit employed in LBP with the motives of improving LPB robustness and complement it with other important texture information. The 2P bits captured both the sign differences between the centre and the neighbour gray values and their respective magnitude information. The experiments conducted on the CK and JAFFE datasets using the SVM classifier showed that CLBP outperformed some other feature representation techniques. Another variant of LBP called uniform LBP (uLBP) has also been considered for the FER task. [141] stated that uLBP is a suitable and reliable image descriptor because of its fundamental image texture properties, with its high percentage in texture image that encourages considerable dimensionality reduction without losing texture context significance, and its tendency to ensure statistical robustness by identifying important local texture pattern. [142] extracted feature from the face using uLBP and reduce the high dimensionality of feature data, utilising the firefly and Great-Deluge algorithm to select an optimal representative subset of the extracted feature. The experiment

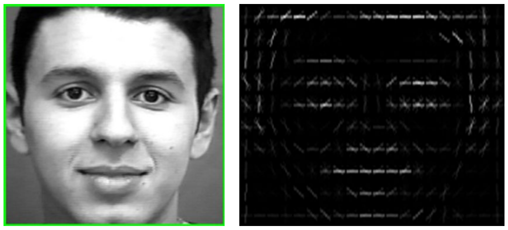


FIGURE 11. The raw image is placed at the left follow by the HOG version of the image.

conducted on the JAFFE dataset using the proposed feature showed that the result produced based on accuracy outperformed the state-of-the-art methods. [141] used Significant Non-Uniform LBP combined with uLBP features to improve the FER recognition rate. He was motivated by the fact that useful micro pattern structural features in facial expression images might be lost if all the non-uniform patterns in the expression image are treated as miscellaneous. He generated features with significant patterns extracted from a non-uniform LBP by considering the transitions from two or more consecutive zeros to two or more consecutive ones, combined with uLBP as FER features.

c: HISTOGRAM OF ORIENTED GRADIENT (HOG)

Dalal *et al.* [143] introduced the histogram of Oriented Gradients (HOG). It is a feature descriptor employed in several fields where objects' characterisation is essential through their shapes and appearance. The histogram of oriented gradients descriptor's motivation is that the distribution of intensity gradients can describe local object appearance and shape within an image and corresponding edge directions [144]. HOG is prominent in object detection as a feature descriptor for image region description. HOG transformation of the happy expression image is shown in Figure 11. HOG starts by dividing an image into blocks and further divides each block into cells. The overlapping blocks made the cell a subcell of many blocks, and then the vertical and horizontal gradient is obtained for each cell's pixel. If $G_y(Y, X)$ is the vertical gradient and $G_x(Y, X)$ is the horizontal gradient, then the magnitude of the gradients are obtained as specified in (3).

$$G(X, Y) = \sqrt{G_x(Y, X)^2 + G_y(Y, X)^2} \tag{3}$$

$$\theta(Y, X) = \arctan \frac{G_y(Y, X)}{G_x(Y, Z)} \tag{4}$$

For each cell, HOG is created. The number of bins with the descriptor is the concatenation of these histograms. Since different images may have different contrast, contrast normalisation is necessary to improve performance. This normalisation results in invariance to changes in illumination and shadowing. Another advantage of HOG is attributed to its operation on local cells, making it invariant to geometric and photometric transformations. HOG was initially utilised for pedestrian detection in static images [143]. [145]–[148] employed HOG as feature descriptor in face detection and recognition. Recently, HOG has been one of the promising feature descriptors for FER. [149] used HOG to encode the deformed components from the detected face and then performed system recognition with linear SVM. The facial parts encoded were the eye-brow and the nose-mouth of the JAFFE database. [92] employed HOG descriptor to describe the representative feature vector for a real-time facial expression system, the feature for each of the patches containing cells concatenated, and the resultant feature vector classified with multiclass SVM. Many other works on FER have also considered HOG mostly as the feature descriptor.

d: PRINCIPAL COMPONENT ANALYSIS (PCA)

Image data is a high dimensional data in which deriving a pattern from it is not easy, but PCA can achieve pattern identification and degree of variabilities in data. PCA uses the dependency between variables of high dimensional data and projects it without losing a significant amount of information into a more tractable lower-dimensional version. PCA tends to find an axis system in data, pointing to maximum covariance in the giving data. The reconstruction of image data results in high dimensionality reduction by using only the significant Eigenfaces responsible for apparent variability. [150] is a thorough survey of FER on PCA. Most of the recent studies in emotion detection used PCA for dimension reduction. PCA is used as a global feature by [3], [151] for expression recognition. [150] in a comprehensive study of facial expression with PCA reported from their research that PCA conducted on facial shape information produced a better result and a better method for FER than the PCA uses facial identities. [152] also enhanced the performance of PCA with Singular Value Decomposition (PCA-SVD) to extract unique features, which provided better performance than both ordinary PCA and LBP + Adaboost. PCA has shown an impressive performance in expression recognition when compared with other Appearance-based features. [153] in their experiment on the JAFFE database, examined the performance of PCA and LDA separately with Euclidean distance as the classifier. Their observation showed that PCA outperformed LDA in terms of recognition rate. Liu *et al.* [154] employed PCA to reduce the hybrid feature dimension of a gray pixel value and extracted LBP from active facial patches of the CK+ database. Then softmax regression classified the dimensionally reduced data space into six basic emotion states under the leave-one-out validation technique.

e: SCALE INVARIANT FOURIER TRANSFORM (SIFT)

SIFT is a detection algorithm introduced by [155], it has four main processing steps; scale-space extrema detection, keypoint localisation, keypoint orientation assignment and keypoint description generation. Scale-space extrema detection deals with keypoint detection, which is achieved by Gaussian (DoG) difference by blurring an image using two different scaling parameters at different octaves of the image Gaussian pyramid. The keypoint obtain at the local extrema is by comparing a pixel with its eight neighbours, nine pixels of the scale above it, and nine pixels of the scale below it. Keypoint localisation ensures a better keypoint by removing low-contrast keypoint and edge keypoint. It can be referred to as a keypoint refiner. Keypoint orientation assignment goal is to make the keypoint robust or invariant to image rotation. Keypoint orientation is achieved by assigning orientation to keypoint. The orientation is computed from the orientation histogram's peak created from the gradient magnitude and direction, calculated from the surrounding keypoint location neighbourhood. The last stage is Keypoint descriptor generation. At this stage, a neighbourhood of 16 × 16 blocks around keypoint is divided into a 4 × 4 size of 16 sub-blocks sub-block creates eight bins orientation histogram that produces a vector of 128 bins value to form the required keypoint descriptor.

Barreti *et al.* [156] extracted SIFT descriptor from the depth of face landmark as a feature for the SVM classifier to addressed person independent problems in 3D expression data. [157] approached emotion recognition from non-frontal facial images by generating super-vectors from the extraction of SIFT features and trained with Edergogic Hidden Markov Model (EHMM). The resultant super-vector was finally classified with Linear Discriminant Analysis (LDA). In [158] keypoints descriptors of SIFT was used as Discriminative SIFT (D-SIFT) features for expression recognition. The investigation conducted by [159] is evidence of the discriminative prowess of SIFT, the result of the experiments on three appearance features; SIFT, LBP and HOG, in a multi-view facial expression analysis showed that SIFT had the best performance. The deep learning approach proposed to solve FER in multi-view images challenge takes a matrix of SIFT features extracted from facial landmarks of images as input feature vector [160]. The model was able to characterise the SIFT feature vectors and their respective high-level semantic information using the corresponding relationship. [161] minimised the small FER data challenge in CNN with dense SIFT feature descriptors and reported that the hybrid of CNN and dense SIFT results in a better performance than using either CNN or CNN with SIFT.

Generally, the strength of appearance-based features lies in capturing transient differences in facial characteristics such as furrows, wrinkles, bulges and many more. However, these features are susceptible to illumination changes and variations in image qualities.

2) GEOMETRIC FEATURE

Geometric features are features extracted statistically from facial landmark displacement. The theory behind this approach is that subsets of face components are more pronounced in facial expression analysis. Geometric feature extraction targets geometric information from facial deformation caused by different kinds of expressions. Geometric based approaches for feature extraction use the Active Shape Model (ASM) or Active Appearance Model (AAM) or their variants to track a dense set of facial points. ASM tends to match groups of model points to an image with a statistical model, and AAM matches an object's shape and texture to an image.

a: ACTIVE APPEARANCE MODEL (AAM)

Cootes [162] introduced the AAM model, which as an extension of the ASM model. AAM successfully forms both the shape and texture of an object. It is categorised as a generative, non-linear and parametric model. AAM has vast applications due to its modelling capability to fix any arising complexity likely to result from high dimensional texture representation. There are three main steps in forming AAM models; (i) connection of shape and texture vectors jointly to each AMM in the training set; (ii) Correlation coefficient matrix computed for the connected shape and the texture vectors in the training set. (iii) Analysis of the correlation coefficient matrix with PCA for each pattern in the training group.

Both ASM and AAM prove to be relevant in facial affective computing, and their application reported severally in literature [163]–[165]. AMM application to FER is frequent in a sequence or Spatio-temporal data. The system proposed in [166] is a real-time system and extracted independent AAM with the aid of the Inverse Compositional Image Alignment (ICIA) method for expression recognition. [167] enhanced the shape produced from the extraction of AAM with second-order minimisation to mitigate large FER errors, to develop FER robust to real-world challenges [168] extracted AAM from edge images rather than gray images, and the report showed that the system is robust against lighting variation. In the pain detection system proposed in [169], AAM was used to decouple shape feature from appearance feature for proper detection of pain through facial expression analysis. [170] used fuzzy logic to monitor the emotion in the shape and texture feature in the facial expression model with AAM. In [171] AAM served as a detector of fiducial point location on facial expression images at the synthesis of feature extraction in the wild. [172] achieved a system that considered ambiguity in the expression displacement for emotion classification by using AAM for face point specification before applying fuzzy C-means for clustering of the emotions. Geometric features are not affected by the lighting condition, and they are not difficult to register and perform well for some Action Units. Nevertheless, they are not suitable to represent an action unit that does not cause landmark displacement.

3) LEARNED-BASED FEATURE

Learned features are attributed to Artificial Neural networks (ANN) and deep learning. Here, ANN learned the direct representative features from the input without feature extraction mathematical models. [51] use visualisation techniques in deep learning to see the kind of feature that Convolution Neural Network (CNN) is using for classification, they observed that the features at the low level resembled low-level Gabor filters. [173] showed that CNN learned features correspond to Facial Action Units (FACs).

TABLE 6. Feature extraction algorithm summary that include strength, limitation and variants. RIFT (Rotational invariant feature transform), GLOH (Gradient location and orientation histogram).

Feature	Strength	Limitation	Variants
LBP	Simple computation with high discriminating power and invariant to grayscale changes	Affected by image rotation and capture limited structural information	Uniform LBP, LBP rotation invariance, Rotated LBP, and Complete LBP
Gabor Filter	maximize spatial and frequency information and possess high sensitivity to small changes in direction	Computationally intensive, and susceptible to high dimensional complexity	Garbor wavelet and log polar Gabor filter
HOG	Capacity to provide global information in large scales and fine-grained details in small scales. invariant to illumination changes and photometric transformation	Extraction takes more time as final descriptor vector grows larger.	Circular HOG, Rectangular HOG
SIFT	invariant to affine rotation and illumination changes	Affected by image rotation, computationally intensive, susceptible to high dimensional complexity	RIFT, PCA-SIFT, GLOH, Gauss-SIFT
AAM	Efficient speed of computation and tendency to reduce high dimensional complexity, it is robust against lighting condition	Not effective for emotion when their is no obvious change in facial landmark displacement, suffers from generalization problem	Locality Constrained AAM, combination with Appearance-based features.
Learned	Efficient descriptor	computaionally intensive, require high computing resources and large volume of data	possible Neural Networks
Hybrid	Feature compliments each other.	Prone to computational complexity	possible combinations of Appearance and/Geometric features

However, the major problem with applying learned-based features to FER is the lack of sufficient data for a network to learn, resulting in overfitting. Another high performing learning-based feature technique is transfer learning, and transfer learning is very efficient in FER where there is limited data for model training. [174]. Apart from CNN based transfer learning, [175] employed an inductive boosting based transfer learning approach to implementing a person-specific model for AUs detection and pain recognition and aimed at achieving generalisation with available minimum data. Learned features have shown promising results, especially in FER, because of its robustness to illumination, rotation, translation, and head pose challenges.

4) HYBRID-BASED FEATURE

Hybrid features give room for the research question of how best to combine features to achieve ultimate performance. [176] proposed an algorithm that fused LBP and HOG features extracted from CK+ and JAFFE database and reduced the extracted features dimension with PCA after permutated the fusion on several classifiers. He found that the fused features on the softmax classifier produced 98.3% on CK+ and 90% on the JAFFE database. The result is evidence that proper hybrid features could significantly improve the system. [177], in their investigation on the best combination of features for optimum performance of the FER system, discovered that the combination of SIFT and geometric features gave better performance compared to either of the features. Also, the experiment showed that LBP and Gabor filter is better in their combination. Table 6 is the concise information of hand-crafted feature extraction algorithms discussed in this work.

C. MACHINE LEARNING MODELS

The feature classification phase ensures the arrangement of features into their respective classes. Classification or regression is achieved chiefly with machine learning classifiers like; Adaboost, SVM, Artificial Neural Network (ANN), deep learning models, or machine learning regression algorithms like Support Vector Regression, Linear Regression and Regression Tree. This section will consider only the popularly used algorithms for classification in FER.

1) SUPPORT VECTOR MACHINE (SVM)

SVM was introduced by [178] as a supervised learning algorithm. It is a binary classifier to find a separating hyperplane of the maximal distance between two trained support vectors.

$$f(x) = W \cdot x + b \tag{5}$$

W in (5) is given as $\mathrm{W} = \sum_i \alpha_i t_\mathrm{i} x_\mathrm{i}$

SVM kernel was modified and adopted for solving the multiclass problem, Figure 12. shows how multiclass SVM is applied to classified six basic emotions. The application of SVM to multiclass tasks is in two categories; direct and indirect. Direct multi-class SVM is discussed in [179]. [180] used one single optimisation process to distinguish all classes. This approach is possible by designing one objective function for training all K-binary SVMs simultaneously and maximise the margins from each category to the remaining levels. Other multi-class SVM direct approaches include; Simplified Multi-class SVM (SimMSVM) [181], Crammer and Singer's multi-class SVM [182]. The dominant indirect multiclass SVM approach is one versus one and one versus rest. In one Versus one, all possible pairwise classifiers are evaluated and therefore induces K(K-1) individual binary classifier [181]. A new feature is applied to each classifier and categorised them using the classifier with the highest vote. K's separate binary classifiers for K class classification are constructed in the one versus rest SVM multiclass approach. This is possible by first training a classifier using the samples from

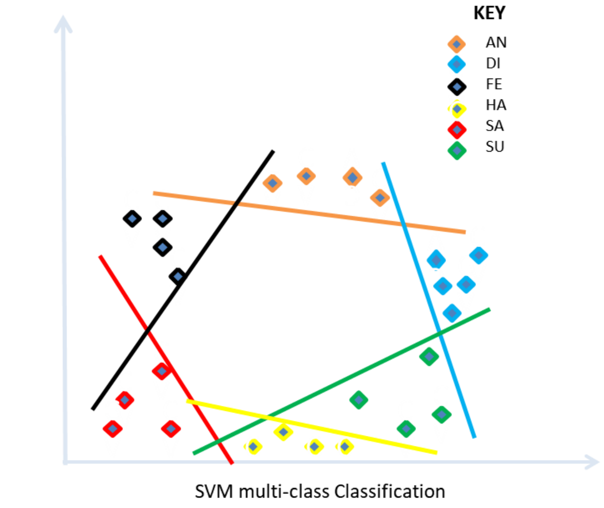


FIGURE 12. **Classification of FER with Multi-class SVM.**

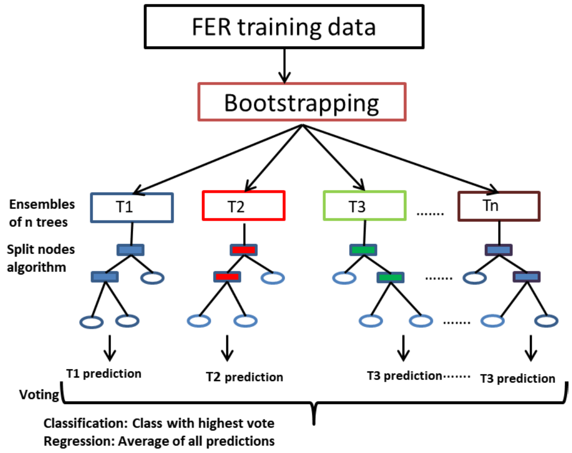


FIGURE 13. **Description of how random forest classifies basic emotions.**

the class as positive samples and regards others as negative. There is an iteration of the process until all the classes have their classifier. SVM is characterised with high performance in terms of accuracy and data size flexibility, and it has proved to be successful in recognising facial expression, based on its generality, more often when the labels are adequately defined [183]. SVM is mostly employed at the classification phase of FER [184] reported that PCA and SVM give better performance on both JAFFE and MUFE databases to individual performances of LPB and PCA. [175] showed that the system achieved appreciable performance when SVM was used to classified boosted geometric features. SVM has also been employed in micro and macro feature classification [50]. It proved so efficient at recognising Facial expressions in real-time [185], [186].

2) ADAPTIVE BOOSTING (ADABOOST)

Adaboost was coined from Adaptive boosting, a boosting algorithm introduced in 1996 by [187]. It builds a robust classifier from a week classifier that vaguely performed better than random guessing. Adaptive boosting is a development over the existing boosting algorithms, and the word boosting came from adapting the new weak classifier to the misclassified data by the previous weak classifier [188]. The robust classifier constructed is a linear combination of weak classifiers. The design of AdaBoost was initially for binary classification problems but recently modified and adapted to various multiclass tasks like FER.

Multiclass Adaboost is achievable by boosting a multiclass classifier. For instance, Allwein *et al.* [189], and Benbouzid *et al.* [190] developed Adaboost with Multi-class Hamming loss (Adaboost.MH). [191] implement Adaboost hypothesis Margin (Adaboost.HM) with the aid of ANN. Also, [192] proposed Adaboost with Binary Decision Tree (Adaboost.BDT) for a multiclass task. SAMME (Stagewise Adaptive Modeling Using Multi-class Exponential Loss Function) was proposed in [193], this version resembles the binary version in a combination of a weak classifier, here the combination was a success by using the stage-wisely forward fitting adaptive model for Multi-class Adaboost. In FER, AdaBoost has been employed as a feature selection and as a classifier. [194] combined AdaBoost with the LBP feature to select the most representative feature for FER called AdaboostLBP. [192] approach the multiclass challenges of FER by incorporating ensembles of Binary Tree Adaboost (BTA), the experiment conducted by [195] established that a multiclass Adaboost that followed the adoption of Classification and Regression Tree (CART) performed better than SVM and MLP in terms of accuracy and speed of computation.

A similar classifier like AdaBoost is Random Forest. Random Forest was introduced by [196], it is an ensemble of trees with bootstrapping and bagging implementation. Its efficiency, computation speed, scalability and easy implementation made it a favourite for many classification tasks. The random forest has been employed mostly as a classifier for facial expression features; Figure 13 illustrates the application of random forest to FER. Random Forest is recently used to classify the facial expression feature, selected by Extreme Learning Auto-Encoder (ELAE) from a complete doubled-LBP features [197]. [198] proposed an extension of random forest termed Pair-wise Condition Random Forest (PCRF). The modified Random Forest learned Spatio-temporal pattern from the fiducial points and facial expression frames' appearance features. PCRF shows a significant result comparable with the existing methods. [49] introduce a cascade of forests model which learns in layers for emotion classification. The result shows that the proposed deep forest showed promising results in a wild environment with sparsely distributed and unbalanced data. Table 7 contains some conventional machine learning classifiers models and their various performances on different feature extraction algorithms.

Generally, the conventional machine learning models are binary classifiers (linear), and adapting them to a non-linear and high dimensional feature-based task, like FER, is a great challenge. This is the major limitation to the performance of

TABLE 7. Some conventional machine learning models analysis and performance.

Model	Features	Classifier	Data	Performance
PU et al. [199]	AAM	Random Forest	CK+	96.38%
Muzammil et al [200]	LBP + PCA	SVM	JAFFE MUFE	87% 77%
Kumar et al. [92]	HOG	SVM	CK+	95%
Zhang and Zheng [160]	SIFT	CNN	BU-3DFE Multi-Pie	80.1% 85.2%
Lilana et al. [172]	SIFT +AMM	Fuzzy C-means	CK+	80.71%
Kauser et al. [201]	LBP	ANN	CK	95.83%
Verma et al. [133]	Gabor	ANN	JAFFE	85.7%
Harit et al. [135]	Gabor Gabor	ANN	JAFFE Yale	81% 57%
Elmadhoun et al. [142]	Gabor	SVM	JAFFE	97.6%
Kumar et al. [92]	HOG	SVM	CK+	95%
Ben et al. [202]	LBP	SVM	MMI CK+ JAFFE	73.3% 97.3 86.7%
Ekweariri et al. [4]	LBP+ Feature selection	NA	BU-3DFE	55%
Wang et al. [203]	CNN feature	Random Forest	JAFFE CK+ FER2013 RAFDB	98.9% 99.9% 84.3% 92.3%

traditional machine learning algorithms applied to FER. Also, conventional machine learning models are shallow learners, and their performance depends on the feature extraction models' output. Nevertheless, research still opens to find the appropriate way of incorporating them into the state-of-the-art method.

The table comprises the model evaluation of some traditional machine learning algorithms for FER.

a: DEEP LEARNING MODELS

Deep learning contains some algorithms which are stacked in a hierarchy of increasing complexity and abstraction. Each of the algorithms applies a non-linear transformation to its input and then uses what it learns to create a statistical model as output. This process is iterative until a detectable level of accuracy is reached. The popularly used deep learning Neural Networks in computer vision is the Convolutional Neural Networks (CNN) and the Recursive Neural Networks (RNN). In FER, CNN is used as a supervised classification task, while RNN is used as an unsupervised classification task, especially FER in real-time.

3) CNN

CNN is one of the deep learning algorithms whose concept evolved from the ANN. [204] introduced CNN in 1998. The design of CNN is purposely for image processing and Computer vision. CNN performs an end to end learning, and the procedure executes in a hierarchy of layers, as shown in Figure 14. Each CNN layer produces representative features ranging from low-level features of the image to a more abstract concept. The process at which CNN automatically learns its representative features emulates the vision mechanism of an animal. That is, the animal visual cortex inspires CNN architectural design. CNN models are self-sufficient in extracting their representative features; there is no need for any pre-calculated features extraction methods. Its high performance contributes immensely to its popularity. The main components of CNN architecture include; convolution layer, pooling layer, dense layer, and fully connected layer.

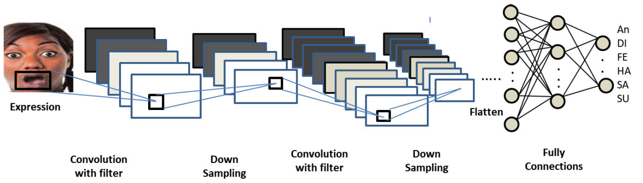


FIGURE 14. A convolution neural network architecture.

4) COMMON CNN ARCHITECTURES

There are quite some impressive number of convolution architectures which have contributed immensely to the field of computer vision, few of the networks include LeNet [204], GoogLeNet [205], ResNet [206], ZFNet [207], VGGNet [208] and AlexNet [209]. Most of the listed networks have been used as a deep base network for the training and classifying facial expression images into basic emotion classes. [174], employed GoogLeNet [205] as the deep base network with a different weight learning algorithm called Peak Gradient Suppression (PGS) for backpropagation. The PGS's essence is to strictly bring the feature representation of non-peak expression closer to their corresponding peak expression. CNN networks complexity varies with the increase in the number of the network components or parameters; this came with the belief that the deeper the network, the better the learning of the data's characteristic features, which improves the network's classification power. This capability makes CNN the most relevant tool in both the machine learning and AI world. Many of the networks are useful for the FER task. Most notably in transfer learning, where expression representative features are learned from a pre-trained network, to improvise for insufficient data challenge in FER. Data insufficiency is the major challenge of employing deep learning to FER tasks because; most of the benchmark datasets are just in their few hundreds or unit of thousands. Table 8 presents a summary of CNN deep architectures.

Application of CNN to FER continues to increase favourably with technology evolution and reducing CNN limitations to FER. Many works have been conducted on FER using CNN as a base classifier in different forms. [210] proposed an Action Unit based deep learning network called AU-inspired Deep Network (AUDN). The CNN network has three phases. The first phase employed the convolution and the pooling operations that learned the representative features called Micro-Action-Pattern. The learned features are to contain information about the local appearance variation. The correlated learned features adaptively combined in the receptive field, which is the second phase. The third phase formed higher-level representations by constructing group-wise sub-networks by applying a multilayer learning process to each receptive field. [211] considered enhancing CNN feature learning capability with some pre-processing procedures so that the network could cope with insufficient data and maximise generalisation capacity. They reported that the system gave an optimal result compare with the state-of-the-art method. [212] proposed an implicit method of ensemble diversity for CNN. They generate different classifiers from a single classifier using parameter variation and fusion of the base classifiers' output. In this case, the classifier considered was CNN. The base classifiers independent CNNs are formed from a random selection of parameters and random selection of CNN architecture, the output generated by each of the base classifiers are fused using the probability-based fusion method. [52] argued that most of the research that automates facial expression considered only strong expressions while weak expressions were left out. The authors presented a CNN network called Deeper Cascaded Peak-piloted Network (DCPN) to join the few. The network design is a version of PPCN by [174], but instead of using GoogLeNet for Network pre-training and fine-tuning, a hybrid of inceptions network called Inception-w, which is a deeper CNN was designed along with a cascaded fine-tuning method used for Pre-training and fine-tuning.

5) RECURRENT NEURAL NETWORK

RNN is a form of Feedforward Neural Networks (FNN) with hidden nodes of memory. The term recurrent emanates from the mechanism of operation of RNN, in the sense that the output of the current input depends on the results obtained from the processing of the previous input(s), as indicated in Figure 15. The hidden nodes make RNN appropriate for many sequence-related tasks like; joined handwriting, voice and speech recognition, Natural Language Processing (NLP) and video processing. Equation (6) is an expression that the current h_t is a function of previous state h_{t-1} and the current input state X_t.

$$h_t = f(h_{t-1}, x_t) \tag{6}$$

Application of activation function to RNN modify (6) to (7)

$$y_t = tanh(Wh_{t-1} + Vx_t) \tag{7}$$

W is the weight of the previous hidden state, V is the weight of the current input state, and tanh is the activation function for non-linearity implementation. The output of RNN is expressed in (7), where y_t is the output state and W is the weight of the output state.

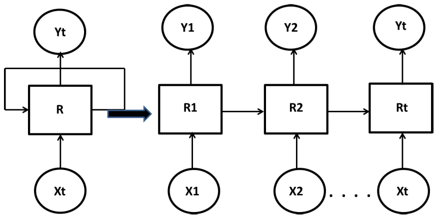


FIGURE 15. **A recursive operation of recurrent neural network.**

Application of FER to a dynamic or a Spatio-temporal environment is possible with the introduction of RNN in [222]–[224]. Nevertheless, the main challenge with RNN is gradient vanishing and exploding. [225] used IRNN (Identity Recurrent Neural Network) proposed in [226] that incorporated ReLus as activation function and Identity matrix as an initialiser to resolve gradient vanishing problem for learning video level representation and classification model in emotion detection in video. Most of the works that modelled FER in a Spatio-temporal environment used LSTM, a modified RNN that remembers past data in memory and overcame gradient vanishing problems. For instance, [53] proposed a model that used ConvLSTM to learn global features for emotion characterisation from the local features generated by 3D-CNN in a spatiotemporal environment. Likewise, in [227], a nested LSTM (T-LSTM and C-LSTM) generated a multilevel feature model from the collection of Spatio-temporal features produced by 3D-CNN for expression characterisation. T-LSTM is a stack of LSTM units purposely designed for temporal dynamics modelling of facial expression, and C-LSTM used the output of T-LSTM to generate the multilevel target features.

Apart from CNN and RNN, other forms of the deep networks also showed commendable performance in their application to facial affective computation. The groups include; Cascaded Networks, Multitask Networks and Generative Adversarial Network (GAN). In cascaded networks, different modules with different functions are sequentially stacked together in hierarchies of dependency. [228] stacked a module for Local Translation Invariant (LTI) using a Multiscale Contraction Convolution Network (MCCN)stacked with Autoencoder that eventually completes the classification task having distinct emotion features from other latent features such as pose and person identity. Similarly, [229] proposed a cascade of DBN and Autoencoder, whereby expression images were trained with DBN to detect the expression region in the face. The output of DBN becomes the input of Autoencoder for expression classification.

Researchers also engage the capability of the GAN network to propose a robust FER model. The strength of GAN is channelled towards removing variation caused by pose and person identity. [55], [230] develop a pose invariant GAN-based network, while [54], [231] works centred on person identity invariant GAN-based model called IA-GAN (Identity Adaptive-GAN) and PPRL-VGAN (Privacy-Preserving Representation learning-VariationalGAN) respectively. Another group of network types is multitasked networks. The motive

TABLE 8. Summary of popular deep convolution neural networks. FER APP: Application of the model to FER, LP: Model learning parameters.

Network	Author and Year	No of Layers	LP	FER APP.	Contribution	Drawback	Network Variants
LeNet	LeCum et al. [204]	5	51,050	[213] [214]	Form the base model for other deep networks	shallow networks	NA
AlexNet	Krizhevsky et al. [209]	8 (5Conv, 3FC)	60Million	[215]	Implement reLu at the convolution layers which aid the speed of computation and also avoid the propagation of negative value into the network.	loss of detail information due to the size of the kernel used	NA
ResNet-50	He and Zhang [206]	50	26 Millions	[216], [217]	Solve saturation problem incur by network depth increase. It adopts batch normalization and skip connection techniques.	Much time of computation is require.	ResNet-34, ResNet-101, ResNet-18.
GoogLeNet	Szegedy et al. [205]	22(9 inception modules, 4Conv, 5FC, 3 softmax)	7Million	Zhou et al. [52], [174]	The introduction of Inception module leads to great reduction of parameters and implemetation of average pooling subsampling techniques eliminates redundant parameters. It resolve overfitting in deep networks. It minimises computational complexity, and also has fast computational speed.	Inception implementation in GoogLeNet almost make the network of no noticeable limitation.	Inception_V3 [218], Inception_V4 and Inception-ResNet [219].
VGGNet	Simoyan et al. [208]	16 (13Conv, 3FC)	138millions	[220], [221]	Able to capture necessary information available with small kernel size. It encourages deeper learning.	computationally complex in time and space. Highly prone to overfitting expecially with small data size	VGG-19, VGGFace

behind multitask networks is to build a robust FER system by creating a network that could identify features that are not relevant and not related to expression so that the network would be able to concentrate only on the relevant information for expression classification. A method proposed in [31], [98] improved FER performance by extending the FER system to include facial landmark localisation. Another example of a FER multitask learning system is Identity Invariant FER introduced in [232] this makes FER robust against subject identity. The method employs two sub-networks (CNN); one of the networks uses expression sensitive loss to learn discriminating expression features, and the other learns discriminating identity features using identity-sensitive loss. The resultant IACNN is robust against subject identity. The work proposed by [233] is a multitask learning called Multi-signal CNN that introduced FER and face verification for network supervision in the FER development system.

Aside from the demand for a large volume of data for CNN to learn discriminating features for its prediction accuracy, other significant CNN challenges include expensive Hyper-parameter tuning. Selecting an adequate number of layers and the components at each layer depends on skills and experience acquired over time. Also, Gradient Vanishing Problem (GVP) is possible due to constant and consistent decreasing in the gradient at each backpropagation operation through multiple levels of non-linearity.

VII. COMPARATIVE STUDY OF FER METHODS

This section provides comparative information based on performance evaluation of some FER methods, categorised into traditional and deep learning methods. The traditional methods are the category of methods that employed handcrafted techniques for feature representation and used machine learning models for classification [128], [201]. While deep learning methods self-learned the representative feature [234], [235]. The study would be based on the experimental results presented in some literature in the field. Table 9 contains the summary of experiments and results of some of FER's traditional and deep learning methods. The experiment conducted by [128] using Compound Local Binary Pattern (CLBP) features and SVM classifier yielded an average accuracy rate of 90% on CK+ data. While the method of [201] using LBP features and ANN classifier give a better recognition rate of 95% also on CK+ data. The traditional methods accuracy performances are high in a controlled environment and very competitive with deep learning methods performances. However, deep learning models gave a recognition rate higher than the traditional methods. The CNN model proposed by [236] reported an average recognition rate of 98% on CK+ data. However, when [160] enhanced the CNN model with SIFT features, they recorded an accuracy of 99.1%. The deep learning model also shows outstanding performance on the JAFFE dataset with the CNN model proposed by [237], which gave an accuracy of 95.8%. Experiments conducted on FER2013, which is a more challenging FER dataset and large, indicate that the work of [238] performed better. [238] combined SIFT features with CNN model to achieve 75.2% accuracy on FER2013. The recognition rate is higher than any traditional methods or pure deep learning predictions on the FER2013 dataset. Experiment on CK+ as sequence dataset shows that Hidden Markov Model (HMM) provided recognition rate of 98.4% [239], which is a good result, but the deep learning model by [240] termed Expression Intensity Invariant Network (EIINet) showed better result with an accuracy of 99.6%. Deep learning Networks have been considered diversely on in-the-wild data and dynamic data. The experiments conducted on AFEW 7.0, the deep model proposed by [241], which hybridised CNN, RNN and c3D for expression recognition in a dynamic environment, provided the state-of-the-art result of 59.02%.

We cannot but also consider some recent experiments, which are graph-based methods discussed in Section V. The methods tend to recover the emotion distribution from

TABLE 9. Summary of some recent experimental results in FER.

Method	Description	Database	No of Classes	Accuracy	Environment	Category
[128]	CLBP+SVM	CK+	6	90.4%	Static	Traditional
[160]	**CNN-SIFT**	CK+	7	**99.1**%	Static	Traditional
[201]	LPB+ANN	CK+	6	95.83%	Static	Traditional
[234]	CNN	CK+	7	97.10%	Static	Deep learning
[235]	GAN	CK+	7	97.30%	Static	Deep learning
[236]	CNN	CK+	6	98.90%	Static	Deep learning
[128]	CLBP+SVM	JAFFE	6	87.5%	Static	Traditional
[237]	**CNN**	JAFFE	7	**95.80**%	Static	Traditional
[242]	LBP+HOG+PCA+SVM	JAFFE	7	94.42%	Static	Traditional
[135]	Gabor Filter + ANN	JAFFE	6	81%	static	Tradtional
[243]	HOG + SVM	JAFFE	7	76.19	Static	Traditional
[244]	GF + KNN	JAFFE	5	68%	statis	Traditional
[244]	HOG + KNN	JAFFE	5	69%	Static	Traditional
[245]	HOG + SVM	JAFFE	6	95.23%	Static	Traditional
[246]	CNN	FER2013	7	75.10%	Static	Deep Learning
[160]	CNN-SIFT	FER2013	7	73.4%	Static	Traditional
[247]	CNN	FER2013	7	75.2%	Static	Deep learning
[238]	**CNN+SVM**	FER2013	7	**75.42**%	Static	Deep Learning
[248]	C3D	CK+	7	91.44%	Sequence	Deep Learning
[239]	HMM	CK+	7	98.54%	Sequence	Traditional
[68]	CRF	CK+	7	93.90%	Sequence	Traditional
[239]	AAM + HMMRF	CK+	7	93.06%	Sequence	Traditional
[249]	Expression Intensity Invariant Network	CK+	7	97.93	Sequence	Deep Learning
[250]	Network Essemble	CK+	6	97.28	Sequence	Deep Learning
[52]	**EIINet**	CK+	7	**99.60**%	Sequence	Deep Learning
[240]	C3D	CK+	7	97.38%	Sequence	Deep Learning
[251]	VGG16-LSTM	AFEW 6.0	7	44.46%	In-the-wild/Dynamic	Deep Learning
[241]	**CNN-RNN-C3D**	AFEW 7.0	7	**59.02**%	In-the-wild/Dynamic	Deep learning
[252]	Cascaded Network	AFEW 7.0	7	47.40%	In-the-wild/Dynamic	Deep Learning
[253]	C3D	AFEW 7.0	7	48.6%	Dynamic	Deep learning

the logical labels of FER data. The graph base models could be semi-supervised (label propagation) or unsupervised (manifold learning). Although these methods are yet to be wildly explored, the manifold feature proposed in [75] with CNN backend gave an accuracy of 76.25% on static and posed data and 66.64% on in-the-wild data. Likewise, the Deep Bi-Manifold CNN (DBM-CNN) model proposed by [34] gave 96.46% on CK+, which is a competitive result in the field.

The high accuracy recorded for the traditional methods could be attributed to the data size. Traditional methods are very efficient in a static environment and with a small data size. In a more challenging environment with a large data size, traditional methods tend to degrade in performance. Although deep learning methods also give high performance, but perform better when there are enough data for the model to learn the representative feature. The more the data, the better the deep learning performance. The combination of deep learning (CNN) and SVM [238] also produced an encouraging performance. The choice of method for a FER task depends on the available data size, type of data (sequence, static, or dynamic), and computational resources' availability. Nevertheless, Deep learning is state-of-the-art because of its universal performance. Its performance with static [235], [236], sequence [52], in-the-wild data and dynamic data [241] is evident in Table 9. Moreover, its challenges with small data size has been alleviated with some optimisation algorithms like; pretrained networks, transfer learning, and the availability of high computing resources.

VIII. DISCUSSION

FER applications still have no limit; They keep evolving with technology. Emotion Recognition and intensity estimation are the significant areas of FER research focus, just as illustrated in Figure 2. The success in detecting AUs' combination from facial expression contributes to compound emotion recognition from facial expression images.

Research outputs in emotion recognition cannot be overemphasised. Facial emotion recognition challenge is an SLL problem. Here, the research goal is a robust model that could tag a basic emotion to a facial expression image. The early works embraced the traditional methods of combining handcrafted feature models and the conventional machine learning models. These models have been diversely considered in different combinations to achieve an optimal result.

Furthermore, the introduction of deep learning models, and the availability of resources that mitigate its application to FER, encourage more research outputs and successes in the field. Deep Learning is still the trending and the state-of-the-art approach to FER. Many methods have been deployed recently to enhance deep learning performance for FER. They include; Enhancement by using a combination of handcrafted features with deep learning feature [254], enhancement by employing a machine learning classifier like decision tree, forest tree and SVM at the output layer of deep learning model [255], [256], Network cascading, use of generative networks, application of some optimisation techniques and others. Deep learning model enhancement for FER is still open research in the field. Recently, the SLL approach to FER has been challenged. The challenge considers that facial expression often reveals more than a single emotion at every display. The argument undermines assigning a logical label to facial expression in the SLL approach models. Using logical labels also denied the FER system of assessing the possible intensity information available in an expression image. Likewise, logical labels prevent models in SLL to consider the correlation among labels, label ambiguity and label inconsistency that are inevitably present in the FER datasets.

FER as regards Intensity estimation has been well studied and also gained noticeable attention in the field. Expression intensity estimation began when some sequence datasets respectively captured the intensity of emotion along with the emotion displayed. Virtually all the studies that considered emotion intensity estimation relied either on annotated sequence datasets or Spatio-temporal data. The analogy that emotion rises from face neutral position to the ON-set and continues to the PEAK before eventually dies as OFF-set is the modality used by many researchers to estimate emotion intensity. The approaches employed in the literature include; Distance-based, Cluster-based, Regression-based and Graphical-based. These methods assign numeric value as the intensity estimation of emotion. This process has been discredited [257], [258] because human intuition does not assign numeric value as a measure of emotional intensity. The only reported ordinal intensity estimation is our model [69], we considered FER and intensity estimation as a multilabel learning task and presented a deep multilabel model, which adequately predicts the emotion and its intensity concurrently, using ordinal metrics.

FER definition as a multilabel task addresses the ambiguity problem in the SLL approach to FER. Adopting a multilabel approach will encourage analysis and recognition of both compound and mixture emotions from facial expressions. Nevertheless, multilabel methods fail to provide information about the proportion of the recognised emotions, and also, emotion intensities are not considered. The multilabel approach to FER is still at the early stage in the field.

Modelling FER as an LDL task efficiently and conveniently resolves label ambiguity, label inconsistency, and correlation among labels in FER databases. Direct application of LDL is achieved in emotion distribution learning [76], [77] model, but direct application of LDL to FER is only possible in datasets with distribution labels. Most of the publicly available FER databases contain logical labels. This limitation is further resolved by label enhancement techniques using clustering [78] and graphical-based methods [78], [80], [223]. The label enhancement techniques encourage more LDL models to explore FER with appreciable results.

The MLL and LDL approaches are yet to gain more attention, unlike the SLL approach, which has been studied differently on static datasets, sequence datasets, spatiotemporal or Video data in controlled or uncontrolled environments.

The available databases for FER research are static, sequence, or Spatio-temporal databases collected in controlled or uncontrolled environments. FER's research using the databases provides promising results, but the results degrade in performance in the real world. This challenge leads to the creation of emotion in the wild databases, possibly collected via internet resources and annotated by experts or using some annotated expert software [48], [56]. Another challenge posed to FER is the unavailability of the FER database in large quantities. Deep learning, the state-of-the-art method in the field, needs a large volume of data to learn the deformation in the face caused by the subtle expression for a reliable prediction. Apart from data size, FER databases also need to consider diversity in cultures, races, age, gender, and degree of emotion intensity at collection and annotation. Also, creating FER datasets with consideration given to correlation among labels in data annotation is highly important for developing an efficient FER system.

A. UNRESOLVED FER CHALLENGES

Despite the achievement in FER, FER research still opens up some unresolved issues. There is a need for a FER robust against the long-existing challenges like; non-frontal head poses, light variation in expression images, data morphology and occlusion. Also, a search in the field is required for optimal ways of combining handcrafted features for FER tasks to achieve better performance. Multi-modal affect recognition is of high interest in the field. Multi-modal suggests how to enhance the FER task with some other affective components (Verbal or non-verbal). Data generability is another obvious challenge in the field; there is a need to explore domain adaptation techniques to ensure cross-database generability. FER applications are yet to explore, despite their broad areas of application. Also, identity specificity, which causes an influx of a person's identity information into different classes that leads to wide intraclass variation and small interclass variation, demands attention. FER database creation and annotations that give preferences to the label correlation and inconsistencies need thorough attention too.

IX. CONCLUSION

We have successfully presented a holistic review of FER that covers its possible research trends based on the machine learning approaches. FER as SLL is the most studied aspect, which is still trending in the field. The MLL and LDL approaches are just gaining attention. It suffices to indicate that both SLL and MLL are possible LDL instances; it is just a matter of threshold definition. Our discussion about some popularly employed models ranging from handcrafted feature models, conventional machine learning models to deep learning models identifies deep learning as the state-of-the-art method and discusses its enhancement with traditional methods. We itemise the unresolved issues in FER together with some future research focus.

REFERENCES

[1] P. Ekman and W. V. Friesen, *Unmaskinh Face a Guide to Recognising Emotion From Facial Clue*. Malor Books, 2003.

[2] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.

[3] A. Dauda and N. Bhoi, "Facial expression recognition using PCA & distance classifier," *Int. J. Sci. Eng. Res.*, vol. 5, no. 5, pp. 570–573, 2014.

[4] A. N. Ekweariri and K. Yurtkan, "Facial expression recognition using enhanced local binary patterns," in *Proc. 9th Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Girne, Cyprus, Sep. 2017, pp. 43–47.

[5] C. Darwin, *Expression of the Emotions in Man and Animals*, 2nd ed., F. Darwin, Ed. New York, NY, USA: Cambridge Univ. Press 2009.

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[7] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[8] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," 2019, *arXiv:1901.02884*. [Online]. Available: http://arxiv.org/abs/1901.02884

[9] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[10] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," *Proc. Comput. Sci.*, vol. 175, pp. 689–694, Jan. 2020.

[11] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. Wróbel, "Emotion recognition and its applications," in *Proc. Adv. Intell. Syst. Comput.*, vol. 300, 2014, pp. 51–62.

[12] Z. Sheng, L. Zhu-Ying, and D. Wan-Xin, "The model of E-learning based on affective computing," in *Proc. 3rd Int. Conf. Adv. Comput. Theory Eng. (ICACTE)*, vol. 3, Aug. 2010, pp. V3-269–V3-272.

[13] C. L. Lisetti and D. J. Schiano, "Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect," *Pragmatics Cognition Pragmatics Cognition*, vol. 8, no. 1, pp. 185–235, May 2000.

[14] C. Yang, A. Qi, H. Yu, X. Guan, J. Wang, N. Liu, T. Zhang, H. Li, H. Zhou, J. Zhu, N. Huang, Y. Tang, and Z. Lu, "Different levels of facial expression recognition in patients with first-episode schizophrenia: A functional MRI study," *Gen. Psychiatry*, vol. 31, no. 2, pp. 1–6, 2018.

[15] B. H. Stamm, "Clinical applications of telehealth in mental health care," *Prof. Psychol., Res. Pract.*, vol. 29, no. 6, pp. 536–542, 1998.

[16] S. Poria, A. Mondal, and P. Mukhopadhyay, "Evaluation of the intricacies of emotional facial expression of psychiatric patients using computational models," Tech. Rep., 2015, pp. 1–286.

[17] D. Joachim and I. Song, "Mental health informatics: Current approaches," *Stud. Comput. Intell.*, vol. 491, pp. 247–253, Nov. 2014.

[18] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," 2020, *arXiv:2002.10392*. [Online]. Available: http://arxiv.org/abs/2002.10392

[19] M. A. Butalia, M. Ingle, and P. Kulkarni, "Facial expression recognition for security," *Int. J. Modern Eng. Res.*, vol. 2, no. 4, pp. 1449–1453, 2012.

[20] A. A. A. Al-modwahi, O. Sebetela, L. N. Batleng, B. Parhizkar, and A. H. Lashkari, "Facial expression recognition intelligent security system for real time surveillance," in *Proc. World Congr. Comput. Sci., Comput. Eng., Appl. Comput. (WORLDCOMP)*, 2012, pp. 1–8. [Online]. Available: http://elrond.informatik.tu-freiberg.de/papers/WorldComp2012/CGV2255.pdf

[21] A. M. Barreto, "Application of facial expression studies on the field of marketing," *Emotional Expression, Brain Face*, pp. 163–189, Jun. 2017.

[22] J.-U. Garbas, T. Ruf, M. Unfried, and A. Dieckmann, "Towards robust real-time valence recognition from facial expressions for market research applications," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 570–575.

[23] G. Yolcu, I. Oztel, S. Kazan, C. Oz, and F. Bunyak, "Deep learning-based face analysis system for monitoring customer interest," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 1, pp. 237–248, Jan. 2020, doi: 10.1007/s12652-019-01310-5.

[24] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "The design and development of a lie detection system using facial micro-expressions," in *Proc. 2nd Int. Conf. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Dec. 2012, pp. 33–38.

[25] N. L. Lopez-Duran, K. R. Kuhlman, C. George, and M. Kovacs, "Facial emotion expression recognition by children at familial risk for depression: High-risk boys are oversensitive to sadness," *J. Child Psychol. Psychiatry*, vol. 54, no. 5, pp. 565–574, May 2013.

[26] M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, p. 4270, Dec. 2018.

[27] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 396–401.

[28] R. Verma, C. Davatzikos, J. Loughead, T. Indersmitten, R. Hu, C. Kohler, R. E. Gur, and R. C. Gur, "Quantification of facial expressions using high-dimensional shape transformations," *J. Neurosci. Methods*, vol. 141, no. 1, pp. 61–73, Jan. 2005.

[29] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognit. Lett.*, vol. 92, pp. 25–32, Jun. 2017.

[30] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, USA, May 2002, pp. 2–7.

[31] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 2–7.

[32] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Grenoble, France, Mar. 2000, pp. 46–53.

[33] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, *Bosphorus Database for 3D Face Analysis* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5372. 2008, pp. 47–56.

[34] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019, doi: 10.1007/s11263-018-1131-1.

[35] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.

[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 211–216.

[37] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, Oct. 2014.

[38] G. Anbarjafari, R. Haamer, E. Rusadze, I. Lsi, and S. Escalera, *Review on Emotion Recognition Databases*. Jan. 2017.

[39] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Shanghai, China, Apr. 2013, pp. 1–5.

[40] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition Emotion*, vol. 24, no. 8, pp. 1377–1388, Dec. 2010.

[41] D. Erhan, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014002159

[42] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, Jul./Sep. 2012.

[43] M. Taini, G. Zhao, S. Z. Li, and M. Pietikainen, "Facial expression recognition from near-infrared video sequences," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[44] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *CoRR*, vol. abs/1708.03985, Aug. 2017. [Online]. Available: http://arxiv.org/abs/1708.03985

[45] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: An addition to the MMI facial expression database," Tech. Rep., 2010.

[46] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *CoRR*, vol. abs/1609.06426, May 2016. [Online]. Available: http://arxiv.org/abs/1609.06426

[47] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4DFAB: A large scale 4D database for facial expression analysis and biometric applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.

[48] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.

[49] O. Ekundayo and S. Viriri, "Deep forest approach for facial expression recognition," in *Proc. Int. Workshops (PSIVT)*, Sydney, NSW, Australia, 2019, pp. 149–160.

[50] H. Khalifa, B. Babiker, R. Goebel, and I. Cheng, "Facial expression recognition using SVM classification on mic-macro patterns," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 6–10.

[51] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," 2017, *arXiv:1705.01842*. [Online]. Available: http://arxiv.org/abs/1705.01842

[52] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *Vis. Comput.*, vol. 34, no. 12, pp. 1691–1699, Dec. 2018.

[53] D. A. A. Chanti and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 363–376, Jun. 2018.

[54] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 294–301.

[55] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 265–270.

[56] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[57] K. K. Lee and Y. Xu, "Real-time estimation of facial expression intensity," in *Proc. IEEE Int. Conf. Robot. Autom.*, Taipei, Taiwan, Sep. 2003, pp. 2567–2572.

[58] C. Quan, Y. Qian, and F. Ren, "Dynamic facial expression recognition based on K-order emotional intensity model," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2014, pp. 1164–1168.

[59] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Intensity rank estimation of facial expressions based on a single image," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Manchester, U.K., Oct. 2013, pp. 3152–3157.

[60] H. Nomiya, S. Sakaue, and T. Hochin, "Recognition and intensity estimation of facial expression using ensemble classifiers," *Int. J. Netw. Distrib. Comput.*, vol. 4, no. 4, pp. 203–211, 2016.

[61] J. Wu and S. Xiao, "Quantitative intensity analysis of facial expressions using HMM and linear regression," in *Proc. 13th ACM SIGGRAPH Int. Conf. Virtual-Reality Continuum Appl. Ind. (VRCAI)*, 2014, pp. 247–250.

[62] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3466–3474.

[63] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 64–84, Feb. 2009.

[64] M. Kim and V. Pavlovic, "Hidden conditional ordinal random fields for sequence classification," in *Proc. ECML PKDD*, in Lecture Notes in Computer Science, vol. 6322, 2010, pp. 51–65.

[65] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.

[66] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 28–43, Feb. 2012.

[67] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012.

[68] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015.

[69] O. Ekundayo and S. Viriris, *Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach*. Cham, Switzerland: Springer, 2020.

[70] R. Plutchik, J. M. Bering, and B. Descriptions, *Contents A Phylogenetic Approach to Religious Origins on the Subcortical Sources of Basic Human Emotions and Emergence of a Unified Mind Science*, vol. 8191. 2001.

[71] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1247–1250.

[72] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 13–26, Jan. 2015.

[73] K. Zhao, H. Zhang, M. Dong, J. Guo, Y. Qi, and Y.-Z. Song, "A multi-labelclassification aprroach for facial expression recognition," in *Proc. Vis. Commun. Image Process.*, Kuching, Malaysia, 2013.

[74] K. Zhao, H. Zhang, and J. Guo, "An adaptive group lasso based multi-label regression approach for facial expression analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 1435–1439.

[75] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13981–13990.

[76] C. Xing, X. Geng, and H. Xue, "Logistic boosting regression for label distribution learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4489–4497.

[77] X. Xi, Y. Zhang, X. Hua, S. M. Miran, Y.-B. Zhao, and Z. Luo, "Facial expression distribution prediction based on surface electromyography," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113683. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420305078

[78] N. El Gayar, F. Schwenker, and G. Palm, *A Study of the Robustness of KNN Classifiers Trained Using Soft Labels*. Berlin, Germany: Springer-Verlag, 2006, pp. 67–80.

[79] M.-L. Zhang, Q.-W. Zhang, J.-P. Fang, Y.-K. Li, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 2057–2070, May 2019.

[80] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4465–4470.

[81] N. Xu, Y.-P. Liu, and X. Geng, "Label enhancement for label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1632–1643, Apr. 2021.

[82] X. Jia, X. Zheng, W. Li, C. Zhang, and Z. Li, "Facial emotion distribution learning by exploiting low-rank label correlations locally," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9833–9842.

[83] Z. Zhang, C. Lai, H. Liu, and Y.-F. Li, "Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection," *Neurocomputing*, vol. 409, pp. 341–350, Oct. 2020.

[84] A. Almowallad and V. Sanchez, "Human emotion distribution learning from face images using CNN and LBC features," in *Proc. 8th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2020, pp. 1–6.

[85] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, pp. 1–9.

[86] H. Filali, J. Riffi, A. M. Mahraz, and H. Tairi, "Multiple face detection based on machine learning," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–8.

[87] M. Böhme, M. Haker, K. Riemer, T. Martinetz, and E. Barth, *Face Detection Using a Time-of-Flight Camera* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5742. 2009, pp. 167–176.

[88] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 643–650.

[89] L. Wang and D. Rajan, "A convolutional neural network approach for face identification," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 5325–5334.

[90] D. Luo, G. Wen, D. Li, Y. Hu, and E. Huan, "Deep-learning-based face detection using iterative bounding-box regression," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 24663–24680, Oct. 2018.

[91] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[92] P. Kumar, S. L. Happy, and A. Routray, "A real-time robust facial expression recognition system using HOG features," in *Proc. Int. Conf. Comput., Analytics Secur. Trends (CAST)*, Pune, India, Dec. 2016, pp. 289–293.

[93] B. Johnston and P. D. Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, Dec. 2018.

[94] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016, doi: 10.1016/j.imavis.2015.11.004.

[95] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[96] M. Bodini, "A review of facial landmark extraction in 2D images and videos using deep learning," *Big Data Cognit. Comput.*, vol. 3, no. 1, p. 14, Feb. 2019.

[97] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," 2017, *arXiv:1712.02494*. [Online]. Available: http://arxiv.org/abs/1712.02494

[98] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Proc. Can. Conf. Comput. Robot Vis.*, Montreal, QC, Canada, May 2014, pp. 98–103.

[99] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, May 2017, pp. 790–795.

[100] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *J. Electron. Imag.*, vol. 25, no. 6, Jun. 2016, Art. no. 061407.

[101] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Macau, China, Sep. 2015, pp. 1–5.

[102] M. Shin, M. Kim, and D.-S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *Proc. 25th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2016, pp. 724–729.

[103] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 433–436.

[104] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with preprocessing stage in automatic emotion recognition," *Proc. Comput. Sci.*, vol. 116, pp. 523–529, Jan. 2017, doi: 10.1016/j.procs.2017.10.038.

[105] C.-M. S.-H. K. Lai and M. Sarkis, "A compact deep learning model for robust facial expression recognition," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2956–2960, 2019.

[106] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 553–560.

[107] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: Towards robust emotion recognition in the wild," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 472–478.

[108] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical frontalization of human and animal faces," *Int. J. Comput. Vis.*, vol. 122, no. 2, pp. 270–291, 2017.

[109] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Fully automatic face normalization and single sample face recognition in unconstrained environments," *Expert Syst. Appl.*, vol. 47, pp. 23–34, Apr. 2016, doi: 10.1016/j.eswa.2015.10.047.

[110] H. Gao, H. K. Ekenel, and R. Stiefelhagen, "Combining view-based pose normalization and feature transform for cross-pose face recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Phuket, Thailand, May 2015, pp. 487–492.

[111] J. Yang, C. Liu, and L. Zhang, "Color space normalization: Enhancing the discriminating power of color spaces for face recognition," *Pattern Recognit.*, vol. 43, no. 4, pp. 1454–1466, Apr. 2010, doi: 10.1016/j.patcog.2009.11.014.

[112] W. Deng, J. Hu, Z. Wu, and J. Guo, "Lighting-aware face frontalization for unconstrained face recognition," *Pattern Recognit.*, vol. 68, pp. 260–271, Aug. 2017.

[113] C. Ferrari, G. Lisanti, S. Berretti, and A. D. Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 1047–1052.

[114] N. Mendez, A. L. Bouza, L. Chang, and H. Mendez-Vazquez, "Efficient and effective face frontalization for face recognition in the wild," in *Prog. Pattern Recognit., Image Anal., Comput. Vis., Appl. 22nd Iberoamerican Congr. (CIARP)*, vol. 1, Valparaíso, Chile. Cham, Switzerland: Springer, 2018, pp. 534–541.

[115] Z. Wu and W. Deng, "One-shot deep neural network for pose and illumination normalization face recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, Jul. 2016, p. 6.

[116] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

[117] W. N. I. Al-obaydy and S. A. Suandi, *Automatic Pose Normalization for Open-Set Single-Sample Face Recognition in Video Surveillance*. Springer, 2019.

[118] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[119] X. Zhu, Y. Liu, Z. Qin, and J. Li, *Emotion Classification With Data Augmentation Using Generative Adversarial Networks*. Cham, Switzerland: Springer, 2018, pp. 349–360.

[120] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A Bayesian data augmentation approach for learning deep models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–10.

[121] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré, "Learning to compose domain-specific transformations for data augmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017.

[122] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[123] C. Lin, M. Guo, C. Li, X. Yuan, W. Wu, J. Yan, D. Lin, and W. Ouyang, "Online hyper-parameter learning for auto-augmentation strategy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6579–6588.

[124] B. Zoph and Q. V. Le, "Neural architecture each with reinforcement learning," *Mach. Learn.*, vol. 2017, pp. 1–16, 2017.

[125] E. D. Cubuk, B. Zoph, D. Man, V. Vasudevan, and Q. V. Le, "Learning augmentation strategies from data," in *Proc. Comput. Vis. Pattern Recognit.*, 2019.

[126] D. Ho, E. Liang, I. Stoica, P. Abbeel, and X. Chen, "Population based augmentation: Efficient learning of augmentation policy schedules," in *Proc. Mach. Learn. Res.*, 2019, pp. 2731–2741.

[127] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu, "Population based training of neural networks," 2017, *arXiv:1711.09846*. [Online]. Available: https://arxiv.org/abs/1711.09846

[128] F. Ahmed, H. Bari, and E. Hossain, "Person-independent facial expression recognition based on compound local binary pattern (CLBP)," *Int. Arab J. Inf. Technol.*, vol. 11, no. 2, pp. 195–203, 2014.

[129] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.*, vol. 93, no. 26, pp. 429–441, Jul. 1946.

[130] S. M. Lajevardi and M. Lech, "Averaged Gabor filter features for facial expression recognition," in *Proc. Digit. Image Comput., Techn. Appl.*, 2008, pp. 71–76.

[131] T. Ahsan, T. Jabid, and U.-P. Chong, "Facial expression recognition using local transitional pattern on Gabor filtered facial images," *IETE Tech. Rev.*, vol. 30, no. 1, pp. 47–52, Sep. 2013.

[132] P. Sisodia, A. Verma, and S. Kansal, "Human facial expression recognition using Gabor filter bank with minimum number of feature vectors," *Int. J. Appl. Inf. Syst.*, vol. 5, no. 9, pp. 9–13, Jul. 2013.

[133] K. Verma and A. Khunteta, "Facial expression recognition using Gabor filter and multi-layer artificial neural network," in *Proc. Int. Conf. Inf., Commun., Instrum. Control*, 2017, vol. 24, no. 9, pp. 1–5.

[134] J. Ou, X.-B. Bai, Y. Pei, L. Ma, and W. Liu, "Automatic facial expression recognition using Gabor filter and expression analysis," in *Proc. 2nd Int. Conf. Comput. Modeling Simulation*, Jan. 2010, pp. 215–218. [Online]. Available: http://ieeexplore.ieee.org/document/5421091/

[135] A. Harit, J. C. Joshi, and K. K. Gupta, "Facial emotions recognition using Gabor transform and facial animation parameters with neural networks," in *Proc. IOP Conf., Mater. Sci. Eng.*, 2018, vol. 331, no. 1, Art. no. 012013.

[136] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[137] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognit. Image Anal.*, vol. 17, no. 4, pp. 592–598, 2007.

[138] B. Tejinkar and S. D. Patil, "Local binary pattern based facial expression recognition using support vector machine," *Int. J. Eng. Sci.*, vol. 7, no. 8, pp. 43–49, 2018.

[139] N. Chitra and G. Nijhawan, "Facial expression recognition using local binary pattern and support vector machine," *Int. J. Innovatice Res. Adv. Eng.*, vol. 3, no. 6, pp. 103–108, 2016. [Online]. Available: http://www.ijirae.com/volumes/Vol3/iss6/17.JNAE10099.pdf

[140] G. Panchal and K. N Pushpalatha, "A local binary pattern based facial expression recognition using K-nearest neighbor (KNN) search," *Int. J. Eng. Res.*, vol. V6, no. 5, pp. 525–530, May 2017.

[141] K. S. Reddy, "A new approach for facial expression recognition using non uniform local binary patterns," vol. 7, no. 3, pp. 20–29, 2018.

[142] A. Elmadhoun, U. Kebangsaan Malaysia, M. J. Nordin, and U. K. Malaysia, "Facial expression recognition using uniform local binary pattern with improved firefly feature selection," *ARO Sci. J. Koya Univ.*, vol. 6, no. 1, pp. 23–32, Apr. 2018.

[143] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, USA, Jun. 2005, pp. 886–893.

[144] M. Dahmane and J. Meunier, "Emotion recognition using dynamic grid-based HoG features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Santa Barbara, CA, USA, Mar. 2011, pp. 884–888.

[145] J. K. J. Julina and T. S. Sharmila, "Facial recognition using histogram of gradients and support vector machines," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Chennai, India, Jan. 2017, pp. 3–7.

[146] X.-Y. Li and Z.-X. Lin, "Face recognition based on HOG and fast PCA algorithm," in *Proc. 4th Euro-China Conf. Intell. Data Anal. Appl.* Cham, Switzerland: Springer, 2018, pp. 10–22. [Online]. Available: http://link.springer.com/10.1007/978-3-319-68527-4

[147] N. Rekha and M. Kurian, "Face detection in real time based on HOG," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 3, no. 4, pp. 1345–1352, 2014.

[148] C. Shu, X. Ding, and C. Fang, "Histogram of the oriented gradient for face recognition," *Tsinghua Sci. Technol.*, vol. 16, no. 2, pp. 216–224, Apr. 2011, doi: 10.1016/S1007-0214(11)70032-3.

[149] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition based on facial components detection and HOG features," in *Proc. Int. Workshops Elect. Comput. Eng. Subfields*, İstanbul, Turkey, 2014, pp. 64–69.

[150] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vis. Res.*, vol. 41, no. 9, pp. 1179–1208, Apr. 2001.

[151] A. Garg and V. Choudhary, "Facial expression recognition using principal component analysis," *Int. J. Sci. Res. Eng. Technol.*, vol. 1, no. 4, pp. 39–42, 2012.

[152] A. P. Gosavi and S. R. Khot, "Emotion recognition using principal component analysis with singular value decomposition," in *Proc. Int. Conf. Electron. Commun. Syst. (ICECS)*, Coimbatore, India, Feb. 2014.

[153] Taqdir and J. Kaur, "Facial expression recognition with PCA and LDA," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 6996–6998, 2014.

[154] Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma, "Facial expression recognition with PCA and LBP features extracting from active facial patches," in *Proc. IEEE Int. Conf. Real-time Comput. Robot. (RCAR)*, Angkor Wat, Cambodia, Jun. 2016, pp. 368–373.

[155] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Kerkyra, Greece, Sep. 1999.

[156] S. Berretti, A. D. Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, İstanbul, Turkey, Aug. 2010, pp. 1–4.

[157] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors," in *Proc. IEEE Int. Conf. Multimedia Expo*, Singapore, Jul. 2010, pp. 1202–1207.

[158] H. Soyel and H. Demirel, "Facial expression recognition based on discriminative scale invariant feature transform," *IET Digit. Library*, vol. 46, no. 5, pp. 4–5, 2010.

[159] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 3–8.

[160] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Aug. 2016.

[161] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN–SIFT aggregator," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell.*, vol. 10607. Springer, 2017, pp. 139–149.

[162] T. F. Cootes, G. J. Edwards, and C. J. Taylor, *Active Appearance Models* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 1407. 1998, pp. 484–498.

[163] K.-E. Ko and K.-B. Sim, "Development of a facial emotion recognition method based on combining AAM with DBN," in *Proc. Int. Conf. Cyberworlds*, Singapore, Oct. 2010.

[164] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model," *Signal Process.-Image Commun.*, vol. 19, no. 8, pp. 723–740, Sep. 2004.

[165] A. S. Dhavalikar and R. K. Kulkarni, "Face detection and facial expression recognition system," in *Proc. Int. Conf. Electron. Commun. Syst. (ICECS)*, Coimbatore, India, Feb. 2014, pp. 1–7.

[166] K.-S. Cho, Y.-G. Kim, and Y.-B. Lee, "Real-time expression recognition system using active appearance model and EFM," in *Proc. Int. Conf. Comput. Intell. Secur.*, Guangzhou, China, Nov. 2006, pp. 747–750.

[167] H.-C. Choi and S.-Y. Oh, "Realtime facial expression recognition using active appearance model and multilayer perceptron," in *Proc. SICE-ICASE Int. Joint Conf.*, Busan, South Korea, 2006, pp. 5924–5927.

[168] C. Martin, U. Werner, and H.-M. Gross, "A real-time facial expression recognition system based on active appearance models using gray images and edge images," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Amsterdam, The Netherlands, Sep. 2008, pp. 1–6.

[169] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1788–1796, Nov. 2009, doi: 10.1016/j.imavis.2009.05.007.

[170] Sujono and A. A. S. Gunawan, "Face expression detection on kinect using active appearance model and fuzzy logic," *Proc. Comput. Sci.*, vol. 59, pp. 268–274, Jan. 2015, doi: 10.1016/j.procs.2015.07.558.

[171] F. A. M. da Silva and H. Pedrini, "Geometrical features and active appearance model applied to facial expression recognition," *Int. J. Image Graph.*, vol. 16, no. 4, pp. 1–17, 2016.

[172] D. Y. Liliana, M. R. Widyanto, and T. Basaruddin, "Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Malang, Indonesia, Oct. 2016, pp. 439–445.

[173] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2015, pp. 19–27.

[174] X. Zhao, X. Liang, L. Liu, T. Li, and Y. Han, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9906, 2016, pp. 425–442.

[175] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines," *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.

[176] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. Ren, "Feature fusion of HOG and WLD for facial expression recognition," in *Proc. IEEE/SICE Int. Symp. Syst. Integr.*, Kobe, Japan, Dec. 2013, pp. 227–232.

[177] L. Zhang, D. Tjondronegoro, and V. Chandran, "Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2012, pp. 1027–1032.

[178] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[179] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proc. Esann*, 1999, pp. 219–224.

[180] V. Vapnik, "The support vector method of function estimation," in *Nonlinear Modeling*, J. A. K. Suykens and J. Vandewalle, Eds. Boston, MA, USA: Springer, 1998, ch. 3, pp. 55–85.

[181] Z. Wang and X. Xue, "Multi-class support vector machine," in *Support Vector Machines Applications*, Y. Ma and G. Guo, Eds. Cham, Switzerland: Springer, 2014, ch. 2, pp. 23–49.

[182] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, no. 2, pp. 265–292, Mar. 2001.

[183] S. W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J. F. Conn, "Improved facial expression recognition via uni-hyperplane classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2554–2561.

[184] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, Malatya, Turkey, May 2015, pp. 14–17.

[185] L. Chen, C. Zhou, and L. Shen, "Facial expression recognition based on SVM in E-learning," *IERI Proc.*, vol. 2, pp. 781–787, Jan. 2012.

[186] P. Michel and R. E. Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proc. 5th Int. Conf. Multimodal Interfaces (ICMI)*, 2003, pp. 258–264.

[187] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 1–15.

[188] H. Fleyeh, R. Biswas, and E. Davami, "Traffic sign detection based on AdaBoost color segmentation and SVM classification," in *Proc. IEEE EuroCon*. Zagreb, Croatia, Jul. 2013, pp. 2005–2010.

[189] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, Sep. 2001.

[190] D. Benbouzid, R. Busa-Fekete, N. Casagrande, F.-D. Collin, and B. K. Egl, "MultiBoost: A multi-purpose boosting package," *J. Mach. Learn. Res.*, vol. 13, pp. 549–553, Mar. 2012.

[191] X. Jin, X. Hou, and C.-L. Liu, "Multi-class AdaBoost with hypothesis margin," in *Proc. 20th Int. Conf. Pattern Recognit.*, İstanbul, Turkey, Aug. 2010, pp. 65–68.

[192] H. Fleyeh and E. Davami, "Multiclass AdaBoost based on an ensemble of binary AdaBoosts," *Amer. J. Intell. Syst.*, vol. 3, no. 2, pp. 57–70, 2013.

[193] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class AdaBoost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[194] S. Prabhakar, J. Sharma, and S. Gupta, "Facial expression recognition in video using AdaBoost and SVM," *Int. J. Comput. Appl.*, vol. 104, no. 2, pp. 1–4, Oct. 2014.

[195] C. S. Fahn, M. H. Wu, and C. Y. Kao, "Real-time facial expression recognition in image sequences using an AdaBoost-based multi-classifier," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2009, pp. 8–17.

[196] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 1–122, 2001.

[197] F. Shen, J. Liu, and P. Wu, "Double complete D-LBP with extreme learning machine auto-encoder and cascade forest for facial expression analysis," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1947–1951.

[198] A. Dapogny, K. Bailly, and S. Dubuisson, "Pairwise conditional random forests for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3783–3791.

[199] X. Pu, K. Fan, X. Chen, L. Ji, and Z. Zhou, "Facial expression recognition from image sequences using twofold random forest classifier," *Neurocomputing*, vol. 168, pp. 1173–1180, Nov. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231215006220

[200] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, May 2015, pp. 276–279.

[201] N. Kauser and J. Sharma, "Facial expression recognition using LBP template of facial parts and multilayer neural network," in *Proc. Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Feb. 2017, pp. 445–449.

[202] N. Ben, G. Zhenxing, and G. Bingbing, Tech. Rep., Sep. 2021.

[203] Y. Wang, Y. Li, Y. Song, and X. Rong, "Facial expression recognition based on random forest and convolutional neural network," *Information*, vol. 10, no. 12, p. 375, Nov. 2019. [Online]. Available: https://www.mdpi.com/2078-2489/10/12/375

[204] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[205] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–12.

[206] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[207] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–9.

[208] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[209] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1–9.

[210] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015, doi: 10.1016/j.neucom.2015.02.011.

[211] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316301753

[212] G. Wen, Z. Hou, H. Li, D. Li, and J. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cognit. Comput.*, vol. 9, no. 5, pp. 597–610, Oct. 2017.

[213] G. Wang and J. Gong, "Facial expression recognition based on improved LeNet-5 CNN," in *Proc. Chin. Control Decis. Conf.*, Jun. 2019, pp. 5655–5660.

[214] Y. Li, X.-Z. Li, and M.-Y. Jiang, "Facial expression recognition with cross-connect LeNet-5 network," *Acta Automatica Sinica*, vol. 44, no. 1, pp. 176–182, Jan. 2018.

[215] X. Chen, X. Yang, M. Wang, and J. Zou, "Convolution neural network for automatic facial expression recognition," in *Proc. Int. Conf. Appl. Syst. Innov. (ICASI)*, May 2017, pp. 814–817.

[216] Y. Chen, J. Du, Q. Liu, and B. Zeng, "Robust expression recognition using ResNet with a biologically-plausible activation function," in *Proc. Pacific-Rim Symp. Image Video Technol.*, Jun. 2018, pp. 426–438.

[217] H. Guo and J. Chen, "Dynamic facial expression recognition based on ResNet and LSTM," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 790, Apr. 2020, Art. no. 012145.

[218] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[219] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[220] H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei, and J. Peng, "Facial expression recognition based on VGGNet convolutional neural network," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 4146–4151.

[221] A. Fathallah, L. Abdi, and A. Douik, "Facial expression recognition via deep learning," in *Proc. IEEE/ACS 14th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Los Alamitos, CA, USA, Oct. 2017, pp. 745–750, doi: 10.1109/AICCSA.2017.124.

[222] A. Graves, J. Schmidhuber, C. Mayer, M. Wimmer, and B. Radig, "Facial expression recognition with recurrent neural networks," in *Proc. Int. Workshop Cognition Technocial Syst.*, 2008.

[223] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 939–947, Jan. 2019.

[224] H. Kobayashi and F. Hara, "Dynamic recognition of basic facial expressions by discrete-time recurrent neural network," *Nippon Kikai Gakkai Ronbunshu, C Hen, Trans. Jpn. Soc. Mech. Eng. C*, vol. 62, no. 594, pp. 644–651, 1996.

[225] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 467–474.

[226] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," 2015, *arXiv:1504.00941*. [Online]. Available: http://arxiv.org/abs/1504.00941

[227] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018, doi: 10.1016/j.neucom.2018.07.028.

[228] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 808–822.

[229] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. Int. Conf. Smart Comput.*, Hong Kong, Nov. 2014, pp. 1–5.

[230] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[231] J. Chen, J. Konrad, and P. Ishwar, "VGAN-based image representation learning for privacy-preserving facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1570–1579.

[232] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 559–565.

[233] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[234] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–29.

[235] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.

[236] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, May 2018.

[237] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2015, pp. 1–8.

[238] M. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[239] C. Wu, S. Wang, and Q. Ji, "Multi-instance hidden Markov model for facial expression recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–6.

[240] S. Kumawat, M. Verma, and S. Raman, "LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 207–216.

[241] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Nov. 2016, pp. 445–450.

[242] B. Islam, F. Mahmud, and A. Hossain, "High performance facial expression recognition system using facial region segmentation, fusion of HOG & LBP features and multiclass SVM," in *Proc. 10th Int. Conf. Electr. Comput. Eng. (ICECE)*, Dec. 2018, pp. 42–45.

[243] S. K. Eng, H. Ali, A. Y. Cheah, and Y. F. Chong, "Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 705, Dec. 2019, Art. no. 012031, doi: 10.1088/1757-899x/705/1/012031.

[244] D. B. Vishal, S. C. Devendra, and M. D. Chaudhari, "Use of KNN classifier for emotion recognition based on distance measures," *Int. J. Eng. Adv. Technol.*, vol. 9, 2019.

[245] R. Safa, H. Rafika, and C. S. Ben, "Facial expression recognition system on SVM and HOG techniques," *Int. J. Image Process.*, vol. 15, 2021.

[246] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3631–3639.

[247] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," Dec. 2016, *arXiv:1612.02903*. [Online]. Available: https://arxiv.org/abs/1612.02903

[248] R. Jack, O. Garrod, H. Yu, R. Caldara, and P. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Nat. Acad. Sci. USA*, vol. 109, pp. 4–7241, Apr. 2012.

[249] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," Mar. 2017, *arXiv:1703.07140*. [Online]. Available: https://arxiv.org/abs/1703.07140

[250] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, Mar. 2019.

[251] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-cue fusion for emotion recognition in the wild," Tech. Rep., Dec. 2016, pp. 458–463.

[252] O. Xi, K. Shigenori, G. H. G. Ester, S. Shengmei, D. Wan, M. Huaiping, and H. Dong-Yan, Tech. Rep., 2017.

[253] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," Tech. Rep., 2017.

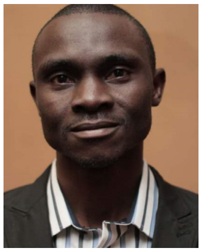
[254] S. M. Lajevardi and M. Lech, "Facial expression recognition using neural networks and log-Gabor filters," in *Proc. Digit. Image Comput., Techn. Appl.*, Canberra, ACT, Australia, 2008, pp. 77–83.

[255] D. M. Vo and T. H. Le, "Deep generic features and SVM for facial expression recognition," in *Proc. 3rd Nat. Found. Sci. Technol. Develop. Conf. Inf. Comput. Sci. (NICS)*, Sep. 2016, pp. 80–84.

[256] A. Ravi, "Pre-trained convolutional neural network features for facial expression recognition," 2018, *arXiv:1812.06387*. [Online]. Available: http://arxiv.org/abs/1812.06387

[257] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," Tech. Rep., 2017, pp. 248–255.

[258] B. Martinez and M. F. Valstar, *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*. Cham, Switzerland: Springer, 2016.



OLUFISAYO S. EKUNDAYO is from Ondo, Nigeria. He received the B.Sc. degree in mathematical sciences (computer science option) from the University of Agriculture, Abeokuta, Nigeria, in 2008, and the M.Sc. degree in computer science from the University of Ibadan, Nigeria, in 2011. He is currently pursuing the Ph.D. degree in computer science with the University of KwaZulu-Natal, South Africa. He worked as a Computer Science Lecturer at Achievers University, Owo, Nigeria, from 2012 to 2018. He has published more than five articles. His research interests include machine learning, pattern recognition, computer vision, and affective computing.

SERESTINA VIRIRI (Senior Member, IEEE) received the B.Sc. degree in mathematics and computer science and the M.Sc. and Ph.D. degrees in computer science, respectively. He has been in academia, since 1998. He is currently a Full Professor in computer science with the University of KwaZulu-Natal, South Africa. He has published extensively in several accredited journals and international and national conference proceedings. He has supervised to completion several Ph.D. and M.Sc. candidates. He is a rated Researcher by the National Research Foundation (NRF) of South Africa. His main research interests include artificial intelligence, computer vision, image processing, machine learning, medical image analysis, pattern recognition, and other image processing related fields, such as biometrics and nuclear medicine. He serves as a reviewer for several machine learning and computer vision-related journals. He also serves on program committees for numerous international and national conferences.

2.2 Conclusion

The chapter presented a comprehensive review that includes facial expression recognition application areas, research trends, and critical analysis of the existing methods. The study revealed the single-label learning approach as the most studied, where facial expression is treated as a multiclass problem. Multilabel and label distribution learning are employed to resolve label ambiguity and data annotation inconsistency problems. Also, existing methods: the handcrafted, machine learning and deep learning models application to facial expression recognition, were critically analysed with their performance and limitations elucidated. Furthermore, the systematic highlight of some unresolved issues in facial expression recognition was carefully presented.

Chapter 3

Facial Expression Recognition and Intensity Estimation

3.1 Introduction

This Chapter presents the frameworks presented and published on facial expression recognition and intensity estimation. The frameworks presented include the deep forest approach for facial expression recognition, a multilabel convolutional neural network for facial expression recognition and ordinal intensity estimation, and facial expression recognition using manifold learning and graph convolutional network.

3.2 Deep Forest Application for Facial Expression Recognition

3.2.1 Introduction

This section presents the investigation of the performance of deep forest on facial expression recognition. Deep forest implementation emulates a layer-by-layer learning feature of the deep neural network to eliminate or reduce computational complexity and present a model that accommodates the small data size available in the field.

3.2.2 Background Model for Deep forest

This section provides detailed information on the machine learning algorithm that underpins deep forest. A deep forest is an ensemble of a random forest learning algorithm.

Random forest is a supervised machine learning technique, which is known for its predictive prowess. Like every ensemble algorithm, random forest builds its predictive capacity from many week learners, and the decision tree is the favourite week learner employs. The algorithm could be used for both classification and regression tasks. Random forest classification tasks will be explored in this work.

Random forest classification task commences from bootstrapping, a technique that generates some data samples called bootstraps from a data distribution. The selected samples are done randomly with replacement of their observation to ensure representativity and independence. Representativity means sampling from the dataset should approximately represent a sampling from real distribution, while independence means bootstraps generated are not correlated. The next step in the classification technique of random forest is Bagging. This process relies on the approximation representative and independence properties of Bootstrapping. It first creates multiple bootstrap samples so that each new bootstrap sample will act as another independent dataset drawn from the new distribution. After that, a week learner (decision tree) is fit to each sample, finds the best feature to split from the sample, and later aggregates them either by hard voting or soft voting, which becomes the output of the ensemble model with low variance. A random forest algorithm is presented in Algorithm 1.

Algorithm 1: Random Forest Algorithm

```
Input: : $X := ( \mathrm{x_1}, \mathrm{y_1}), .... (\mathrm{x_n}, \mathrm{y_n})$
Feature M, N number of trees. Output:  predicted class
H := 0 // intialise tree H
for i := 1 to N do
$k^{(\mathrm{i})} := Random(k, M)$ // random selection of k feature from the entire M feature
$\mathrm{h^i}$ := compute best split point fro k feature // best split point from the bootstrap k is computed
$H := H \cup h^{\mathrm{i}}$ //update tree H
// Performing classification using the tree generated
Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
Calculate the votes for each predicted target.
Consider the high voted predicted target as the final prediction from the random forest algorithm.
```

3.2.3 How Deep Forest Works

A deep forest could be considered a cascade of random forests where each cascade is an ensemble of forest trees. In other words, a deep forest is an ensemble of ensemble of decision trees. The motive behind deep forest is to make shallow learning algorithms like a decision tree to learn in layers - a feature of a deep neural network. A deep forest begins Implementation with a feature extraction mechanism called multi-grained scanning - a feature similar to the convolution process in the deep neural network. The multi-grained scanning process extracts feature information from an image feature by striding filter or window of the desired size over the image feature. If an image is given as I(x,y), and window size is given as W(f,f), then the resultant feature after multi-grained scanning is M presented in (3.1).

The second phase is the deep forest structure, a cascade structure in the form of a progressive nested structure of different forest trees. The model implements four different forest trees classifiers (two Random Forest, ExtraTree and Logistic Regression classifiers), and each of the forest trees contains 500 trees. The difference between the forest trees is their mode of selecting the representing feature for a split at every tree node.

The learning principle of the deep forest is the layer to layer connectivity. A layer communicates with its immediate preceding layer by taking as input the preceding layer's output. The efficiency of the cascade structure lies in its ability to concatenate the original input with the features inherited at each layer. The motive is to update each layer with the original pattern and make the layer achieve reliable predictions. The concatenation of the original input layer thus enhances the generalization of the structure. Forest processes start with bagging (bootstrap Aggregation). If there is an N data sample, then some numbers n subsets of R randomly chosen samples with replacement is created such that each subset is used to train a tree, and the aggregate forest contains n trees. The tree growth for each of the forests starts from the root with the whole dataset, then each node containing an associate sample is split into two with reference to the randomly selected feature from the forest. The two subsets are then distributed on the two children nodes, and the splitting continues until there is a pure sample of a class at the leaf node of the tree or the predefined condition is satisfied. Layer by layer of forest learning is presented in Figure 3.1

$$M = (x - f + 1) * (y - f + 1) \tag{3.1}$$

3.2.4 Mathematical Illustration of the framework

Data description Let $X = \mathbb{R}^m$ represents m dimensional real feature input space having real value components, and let Y = $\{y_1, ...y_n\}$ be the output space, where n is the number of classes and $n \in \mathbb{N}$. Then every sample $x_i \in X$ has corresponding $y_i \in Y$, where $i = 1 \ldots n$ and the training sample Δ_i is:

$$\Delta_i = \{(x_i, y_i)\}_{i=1}^n$$

At each layer there are forests and each forest contains learning algorithms $\{\alpha\}_{i=1}^n$ that could be regarded as functions which give the image of the input data

Algorithm 2: Deep Forest Algorithm

```
Input: : $X := (x_1, y_1), .... (x_n, y_n)$ feature m, N numbers of trees
Output: predicted class
for n := 1 to N // where N is the number of tree and n<=N
for i := 1 to L do where L is the number of layers and i <= L
train all the trees from n forest at i layer
compute P probability vector for every $x_i \in X$
compute $x_i \cup (x_i, P)$
end for
Use $x_i \cup (x_i, P)$as input in next i
end for
terminate while there is no significant improvement in the output
```

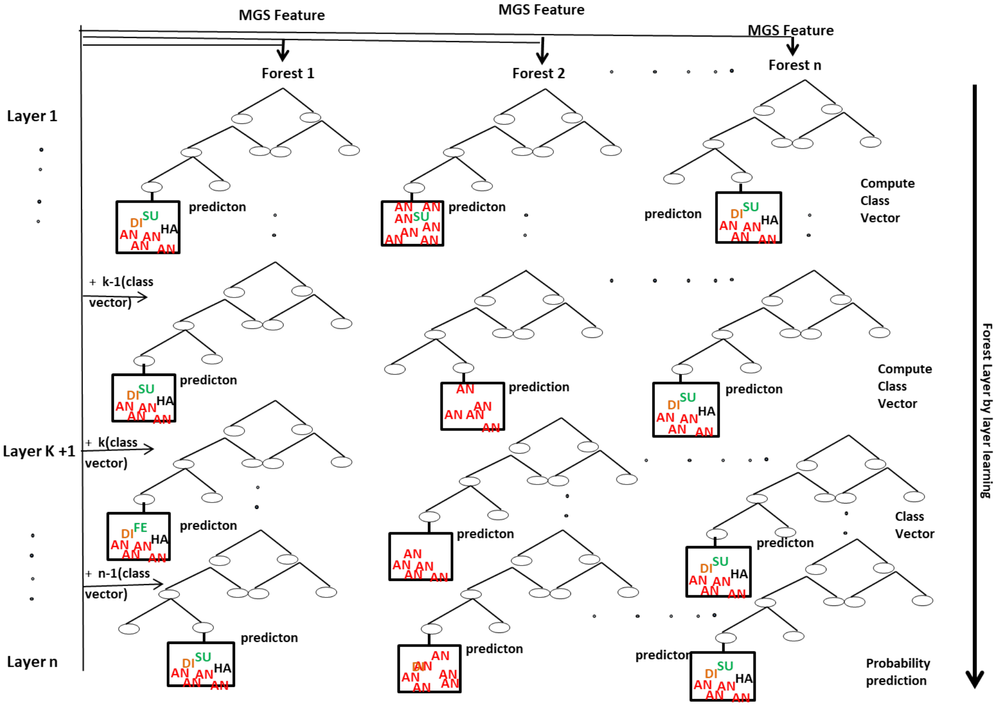


Figure 3.1: Block diagram description of the processes in ML-CNN.

as the output of the forest. Then each forest in the first layer $L_1 \in L$ contains set of learning function say α^1 with general behaviour: $\alpha^1 : X \to X^1$, where X is the input data into the layer 1 and X^1 is the image of X, L represents possible layers and L $= \{L_1, \ldots, L_n\}$ such that $n \in \mathbb{N}$. Then all functions in layer 1 are represented as:

$$\alpha^1 = \alpha_1^1, ..., \alpha_n^1$$
$$X^1 = \alpha_1^1(X), ..., \alpha_n^1(X)$$

this implies that a new data is gotten at the output of L_1, then data sample is updated for the next layer:

$$\Delta_i = \Delta_i^1 = \{(x_i^1, y_i)\}_{i=1}^n$$

The process continues as long as there is a significant performance in the model at every successive layer. At every layer k such that $1 < k < n$ in the model where there is appreciable improvement in the performance of the model, it suffices to recall that the input to layer k is $X^{(k-1)}$

$$X^k = \alpha_1^k(X) \times .... \times \alpha_n^k(X)$$

the output of layer k is:

$$\Delta_i^k = \{(x_i^k, y_i)\}_{i=1}^n \quad \text{for all} \quad n \in \mathbb{N}$$

the layer stop growing at layer n where there is no significant increase in the performance of the model. At layer n there is an assurance of having $\mathrm{x_i}^{\text{(n-1)}}$ converging closely to y_i. Note that, the output of each layer is the average of the probability distribution for instances in the leaf node of the trees for each forest. let $P_t = P_1, \ldots, P_s$ be the class vector probability of each node of the tree. For each sample of input $X^{(n-1)}$ the probability vector of the leaf node is given as:

$$P_t^n(X^{(n-1)}) = (P_1^n(x_i^{(n-1)}), ..., P_s^n(x_i^{(n-1)}))$$

then the output of Forest β in a layer n is the average of the probability vectors of all trees in the forest; given as:

$$\beta_j = \frac{1}{J} \sum_{j=1}^{J} P_j(X_s)$$

where J is the number of trees in a Forest and s is the number of class vector estimation at the leaf node.

This work was presented in [1] and published in [2]

[1] The 9th Pacific-Rim Symposium on Image and Video Technology 18-22 November, 2019,Sydney, Australia.

[2] Ekundayo O., Viriri S. (2020) Deep Learning Approach for Facial Expression Recognition. In: Dabrowski J., Rahman A., Paul M. (eds) Image and Video Technology. PSIVT 2019. Lecture Notes in Computer Science, vol 11994. Springer, Cham.

Deep Forest Approach for Facial Expression Recognition

Olufisayo Ekundayo and Serestina Viriri(✉)

School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, Durban 4000, South Africa
sunfis1979@gmail.com, viriris@ukzn.ac.za

Abstract. Facial Expression Recognition is a prospective area in Computer Vision (CV) and Human-Computer Interaction (HCI), with vast areas of application. The major concept in facial expression recognition is the categorization of facial expression images into six basic emotion states, and this is accompanied with many challenges. Several methods have been explored in search of an optimal solution, in the development of a facial expression recognition system. Presently, Deep Neural Network is the state-of-the-art method in the field with promising results, but it is incapacitated with the volume of data available for Facial Expression Recognition task. Therefore, there is a need for a method with Deep Learning feature and the dynamic ability for both large and small volume of data available in the field. This work is proposing a Deep Forest tree method that implements layer by layer feature of Deep Learning and minimizes overfitting regardless of data size. The experiments conducted on both Cohn Kanade (CK+) and Binghamton University 3D Facial Expression (BU-3DFE) datasets, prove that Deep Forest provides promising results with an impressive reduction in computational time.

Keywords: Facial Expression Recognition · Deep Neural Network · Deep Forest

1 Introduction

Deep forest learning is a recent method initiated by [14,21] with the motive of approaching classification and regression problems by making a conventional classifier (shallow learners) like the random forest (decision tree) to learn deep. The prevalence of Deep Neural Network (DNN) in Machine Learning (ML) and Artificial Intelligence (AI) can never be overemphasised. Deep learning is said to be as old as Artificial Neural Network (ANN) but went into hibernation due to its computational complexity and the demand for the large volume of data [5]. In the recent years, the availability of sophisticated computational resources and invention of the internet that give room for the collection of large datasets play a remarkable role in bringing deep learning back into the forefront of machine learning models. Deep learning has proven its worth in several areas of classification and regression computation with an efficient and optimal solution.

J. J. Dabrowski et al. (Eds.): PSIVT 2019 Workshops, LNCS 11994, pp. 149–161, 2020.
https://doi.org/10.1007/978-3-030-39770-8_12

Beyond reasonable doubt, deep learning outperformed the conventional classifiers in most machine learning tasks like; image processing, computer vision, pattern matching, biometrics, bioinformatics, speech processing and recognition, etc. Nevertheless, despite the computational prowess of Deep learning, its quest for large datasets and computational resources consumption is still a challenge. Therefore, there is a need to explore other machine learning models and see the opportunities to enhance their capability for better efficiency and accuracy.

Deep forest is still very new in machine learning and this implies that its application is yet to be explored. Both Forward Thinking Random Forest and gcForest are the popularly available deep forest models. And the reports of the models give the similar performance even if not more, as DNN in their experiments on MNIST dataset, with additional advantages of low computational time, limited hyper-parameter tuning and dynamic adaptation to the quantity of available dataset. Our task in this paper is to develop a deep learning model from the ensemble of forest trees for the classification of facial expression into six basic emotions, while depending on the forest tree inherent affinity for multi-class problems. Facial expression recognition is a multi-class problem and its goal is to detect human affective state from the deformation experience in the face due to facial muscles response to emotion states. To the best of our knowledge, this work is the first of its kind that engages a layer by layer enssemble of forest tree approach to the task of facial expression classification.

In this paper; Sect. 2 contains the details description of the related works, it captures the performances and the limitations of some of the classification models on facial expression recognition data. Section 3 contains a brief introduction to random forest and the description of the proposed deep forest framework for facial expression recognition. In Sect. 4 we discuss the databases for the experiments while Sect. 5 contains details of the experiment performed and the result analysis. Section 6 is the conclusion of the work.

2 Related Works

The complexity of facial expression and the subtle variations in its transition give rise to several challenges experienced in the field. One of the major challenges of facial expression is its classification into the six category of classes proposed by [9]. Many classifiers and regression algorithms have been proposed severally to address the challenge, the methods include Support Vector Machine (SVM) [15], Boosting Algorithm (AdaBoost) [7], Convolution Neural Network (CNN) [2] Decision Tree [18], Random Forest [4], Artificial Neural Network (ANN) [19], to mention a few. The listed classifiers have reportedly produced various promising results depending on the approach.

The impressive performance of Decision tree towards classification problems makes its evident application in several machine learning fields. [18] used decision tree to classify feature extracted from a distance based feature extraction method. Although there are not many works in facial expression recognition with decision tree method because of overfitting challenge in its performance with high dimensional data [16], the available ones are either presented its boosting (AdaBoost) or an ensemble (forest tree) version. Decision tree has been graciously enhanced by the introduction of Forest tree [3]. A random forest tree is an ensemble of learner algorithms in which individual learner is considered to be a weak learner. Decision tree algorithm has been widely explored as a weak learner for random forest tree, and this is likely the reason for describing a random forest as the ensemble of decision trees. [6] in their work extends the capability of random forest tree to a spatiotemporal environment, where subtle facial expression dynamics is more pronounced. The model conditioned random forest pair-wisely on the expression label of two successive frames whereby the transitional variation of the present expression in the sequence is minimized by the condition on the most recent previous frame. [12] hybridized deep learning and decision tree, and the hybridization was based on the representation learning of deep learning feature and the divide and conquer techniques of decision tree properties. A differentiable backpropagation was employed to enhance the decision tree to achieve an end to end learning, and also preserving representation learning at the lower layers of the network. So that the representation learning would minimize any likely uncertainty that could emerge from split nodes and thus minimized the loss function. The concept of Deep Forest is beyond the integration of decision tree into Deep Neural Network as proposed in [12]. [14,21] thoroughly highlighted; computation complexity cost as a result of using backpropagation for the multilayers training of nonlinear activation function, massive consumption of memory during the training of complex DNN models, overfitting and non-generalization to small volume of data and complexity in hyperparameter tuning; as the challenges encountered while implementing Deep Neural Network. Therefore, there is a need for a deep learning model type that would minimize the challenges in the existing deep learning models. [14] proposed a deep learning model (Forward Thinking Deep Random Forest) different from ANN, in which the neurons were replaced by a shallow classifier. The network of the proposed model was formed by layers of Random Forest, and decision tree which is the building blocks of forest tree was used in place of neurons. The model was made to train layer by layer as opposed to the once-off training complexity and rigidity experienced in DNN. Likewise, the evolving Deep Forest learning (gcForest) proposed by [21] ensures diversity in its architecture, where the architecture consists of layers with different random forests. Both models successfully implement deep learning from Random Forest without backpropagation. Although the mode of achieving this slightly differs, while gcForest ensures connection to the subsequent layer using the output of the random forest of the preceding layer, the connection to the subsequent layer in FTDRF is the output of the decision tree in the random forest of the preceding layer. As earlier stated, it was reported that both models outperform DNN on the performance evaluation experiment on MNIST datasets.

3 Deep Forest Learning

Before providing the details of Deep Forest learning operations, it suffices to discuss the basic concept of Random Forest tree.

3.1 Random Forest

Random Forest tree was introduced by Breiman [3], before the advent of Breiman's work, tree learning algorithm (decision tree) had been in existence, the algorithm was effective and efficient. Its implementation could either be shallow or a full grown tree (deep tree). Shallow tree learning model has a great affinity for overfitting resulting from the model high bias and low variance features, which is often addressed by boosting (AdaBoost) algorithm. Breiman established the ensemble idea on the early works of [1,8,20] and proposed a random forest algorithm which is efficient for both regression and classification tasks. Breiman implements both bootstrapping and bagging techniques by randomly creates several bootstrap samples from a raw data distribution so that each new sample will act as another independent dataset drawn from the true distribution. And after fit, a weak learner (decision tree) to each of the samples created. Lastly, computes the average of the aggregate output. The operation that would be performed on the aggregate of the output of the weak classifiers is determined by the task (classification or regression). In case of a regression problem, the aggregate is the average of all the learners' output and if classification the class with the highest volt is favoured. Random forest is known for its fast and easy implementation, scalable with the volume of information and at the same time maintain sufficient statistical efficiency. It could be adopted for several prediction problems with few parameter tuning, and retaining its accuracy irrespectives of data size.

3.2 Proposed Facial Expression Deep Forest Famework

Deep forest learning architecture as presented in Fig. 1 is a layer by layer architecture in which each layer comprises of many forests, a layer links with its successive layer by passing the output of the forest tree as input to the next layer in the architecture. This work enhance the deep forest model proposed by [21] by introduction of trees with different features at strategic positions for better performance. The model consists of two phases; the feature learning phase and the deep forest learning phase. The feature learning phase is integrated for the purpose of feature extraction similar to convolution operation in DNN. It uses windows of different sizes to scan the raw images (face expression images), in a process of obtaining a class representative vector. The class vector is a N-dimensional feature vector extracted from a member of a class and then use for the training of the Deep Forest.

The second phase is the main deep forest structure; a cascade structure in the form of a progressive nested structure of different forest trees. The model implements four different forest trees classifiers (two Random Forest, ExtraTree and Logistic Regression classifiers), and each of the forest trees contains 500 trees. The difference between the forest trees is in their mode of selecting the representing feature for a split at every node of the tree.

The learning principle of the deep forest is the layer to layer connectivity, that is, a layer communicates with its immediate predecessor layer by taking as input, the forest tree output o f the preceding layer. The efficiency of the cascade structure lies in its ability to concatenate the original input with the features inherited at each layer. The motive is to update each layer with the original pattern and also to make the layer achieves reliable predictions. The concatenation of the original input layer thus enhances the generalization of the structure. Each layer is an ensemble of forests, the connection from one layer to another layer is done through the output of the forests. Forest processes start with bagging (bootstrap Aggregation). If there is N data sample, then some numbers n subsets of R randomly chosen samples with replacement is created such that each subset is used to train a tree, and the aggregate forest contains n trees. The tree growth for each of the forests starts from the root with the whole dataset, then each node containing an associate sample is split into two with reference to the randomly selected feature from the Forest. The two subsets are then distributed on the two children nodes, and the splitting continues until there is a pure sample of a class at the leaf node of the tree or the predefined condition is satisfied.

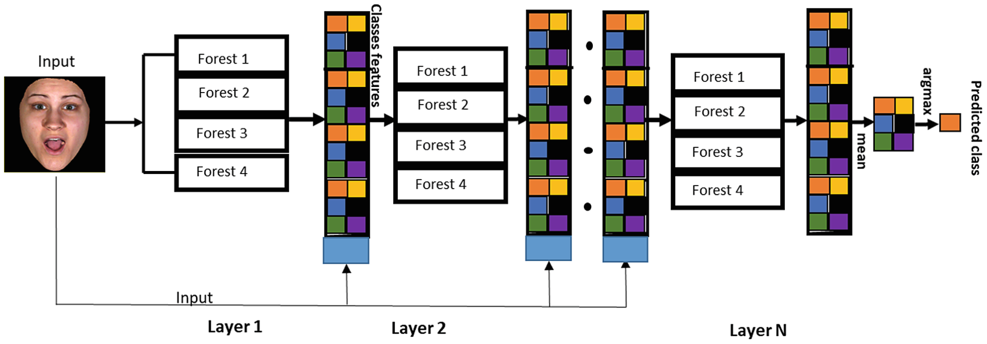


Fig. 1. Deep forest architecture

For each instance of a class, class distribution estimation is computed, and then averaging across all trees for each forest. This becomes the class vector to be concatenated with the original feature vector and send to the cascade next layer as input. Which implies each class will have one class vector, the number of augmented features extracted depends on the number of class multiply by the number of trees in the deep forest model. In order to control overfitting, K-fold is used to generate the class vector for each forest. At every layer expansion, cascade performance evaluation is estimated. At a point in the training where there is no significant improvement in the performance, the training is halt. This account for the control that Deep Forest has over its architecture.

3.3 Mathematical Illustration of the Framework

Data description Let $\chi = R^m$ represent the input space, and let $Y = y_1, ...y_c$ be the output space. Then every sample $x_i \in \chi$ has corresponding $y_i \in Y$ the training sample Δ_m is:

$$\Delta_m = (x_1, y_1), ..., (x_m, y_m)$$

At each layer there are forests and each forest contains learning algorithms that could be regarded as functions which give the image of the input data as the output of the forest. Then each forest in the first layer, L_1 contains set of learning function say α^{l1} with general behaviour: $\alpha^{l1} : \chi_i \rightarrow \chi_i^{l1}$ where χ_i is the input data into the layer1 and χ_i^{l1} is the image of χ_i then all functions in layer1 are represented as:

$$\alpha^{l1} = \alpha_1^{l1}, ..., \alpha_n^{l1}$$
$$\chi_i^{l1} = \alpha_1^{l1}(\chi_i), ...\alpha_n^{l1}(\chi_i)$$

this implies that a new data is gotten at layer 1, which means:

$$\Delta_m = \Delta_m^{l1} = (\chi_1^{l1}, y_1), ..., (\chi_m^{l1}, y_m)$$

The process continues as long as there is a significant performance in the model at every successive layer. At every layer k in the model where tree is appreciable improve in the performance of the model, it suffices to recall that the input to layer k is $\chi_i^{(k-1)}$

$$\chi_i^{lk} = \chi_1^{lk} \times \chi_2^{lk} \times .... \times \chi_n^{lk}$$
$$\chi_i^{lk} = \alpha_1^{lk}(\chi_i), ..., \alpha_n^{lk}(\chi_i)$$

the output of layer k is:

$$\Delta_m^{lk} = (\chi_1^{lk}, y_1), ...(\chi_m^{lk}, y_m)$$

the layer stop growing at layer n where there is no significant increase in performance of the model. At layer n there is an assurance of having $\chi_i^{l(n-1)}$ converging closely to y_i. Note that, the output of each layer is the average of the probability distribution for instances in the leaf node of the trees for each forest. Let $P = p_1, ..., p_t$ be the class vector probability of each node of the tree. For each sample of input $\chi_i^{l(n-1)}$ the probability vector of the leaf node is given as:

$$P_i^{ln}(\chi_i^{l(n-1)}) = (P_1^{ln}(\chi_i^{l(n-1)}), ..., P_t^{ln}(\chi_i^{l(n-1)}))$$

then the output of Forest β in a layer l^n is the average of the probability vectors of all trees in the forest; as given in (1):

$$\beta_j = \frac{1}{J}\sum_{j=1}^{J} P_j(\chi_t) \tag{1}$$

where J is the number of trees in a Forest and T is the number of class vector estimation at the leaf node.

4 Database

In this section we briefly introduce the two databases (BU-3DFE and CK+) that we are proposing for the experiment here. Figures 2 and 3 are the respective samples of the expression images in BU-3DFE and CK+.

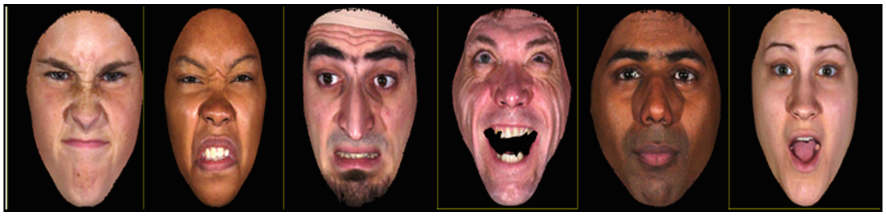


Fig. 2. Selected expression images samples from BU-3DFE datasets. The arrangement from left: Angry, Disgust, Fear, Happy, Sad and Surprise

4.1 Binghamton University 3D Facial Expression (BU-3DFE)

This database was introduced at Binghamton University by [17], it contains 100 subjects with 2500 facial expression models. 56 of the subjects were female and 44 were male, the age group ranges from 18 years to 70 years old, with a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. 3D face scanner was used to capture seven expressions from each subject, in the process, four intensity levels were captured alongside for each of the 6 basic prototypical expressions. Associated with each expression shape model, is a corresponding facial texture image captured at two views (about +45° and −45°). As a result, the database consists of 2,500 two view's texture images and 2,500 geometric shape models.

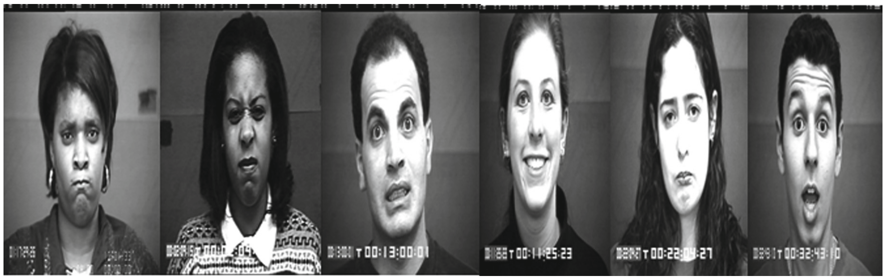


Fig. 3. Selected expression images for each of the emotion states from CK+ datasets. The arrangement from left: Angry, Disgust, Fear, Happy, Sad and Surprise

4.2 Cohn Kanade and Cohn Kanade Extension (CK and CK+) Database

[11] released a facial expression database in 2000, the database contains 97 subjects between the ages of 18 and 30; 65 were female and the remaining 35 were male. The subjects were chosen from multicultural people and races. There were 486 sequences collected from the subjects and each sequence started from neutral expression and ended at the peak of the expression. The peak of the expressions was fully FACS coded and emotion labeled, but the label was not validated. [13] itemized three challenges with CK databases challenges; invalidation of emotion labels because it did not depict what was actually performed. Unavailable common performance metrics for algorithm performance evaluation, as a result of no standard protocol for a common database. [13], having identified the challenges with CK database proposed its extension termed extended Cohn Kanade (CK+) database. In CK+ the number of subjects was increased by 27 and the number of sequence by 22, there were slight changes in the metadata also, age group of the subject ranged between 18 and 50, male was 31, and female was 69. The emotion labels were revised and validated using FACS investigator guide as a reference and confirmed by appropriate expert researchers. Leave-one-out subject cross-validation and area underneath the Receiver Operator Characteristics curve were proposed as metrics for Algorithm performance evaluation.

5 Experiment

The experiment was conducted on two datasets; the Cohn Kanade extension (CK+) and the Binghamton University 3D Facial Expression (BU-3DFE) datasets. We used only the peak images for the six basic emotion states (Anger, Disgust, Fear, Happy, Sad, Surprise) of 2D images from each of the data sets, and the total number of expression images used from BU-3DFE is 600 (100 images per emotion, 54 female and 46 male). In CK+ dataset; the total number of images extracted was 309 but the number of images per emotion varied (AN = 45, DI = 59, FE = 25, HA = 69, SA = 28, SU = 83). We split each of the extracted data into two; the training set (80%) and the validation set (20%). The training set was used to train the forest and the validation set was used for the performance evaluation. The model depth (the number of layers) is automatically determined, each layer consists of three different pairs of forests, and each forest contains 500 trees.

Before feeding the images as input for processing data processing techniques such as face detection, face alignment and histogram equalization were applied on the data so as to minimise data redundancy and intensity variation that may possibly challenge the performance of the system. As earlier stated we split the input into the training data and the validation data. Growing the forests with the training data set, we used 5-fold cross-validation to minimized chances of overfitting.

We tested the trained model on the validation set and passed each instance of the validation as representative feature to the cascade forest classification process. The output of the cascade forest returned probability predictions from each forest in the last layer of the cascade. As a result, the mean of the predictions was computed, and finally, the class with maximum value is the outcome of the prediction. For performance evaluation we use accuracy as our metrics and also employ confusion matrix for proper analysis of the result.

Furthermore, we conducted an investigation on the effect of number of classifiers on the behaviour of Deep Forest model. Initially, on both datasets (CK+ and BU-3DFE) we used 4 forest classifiers, and obtained average accuracy of 93.22% with only 5 layers added and 7 estimators in each layer for CK+ dataset. When each of the classifiers was doubled, the accuracy remained but ten layers were added with 7 estimators in each layer. This is different in the case of BU-3DFE dataset, the initial 4 classifiers gave accuracy of 57.98% and added 8 layers with 8 estimators in each layer. When each of the classifiers was doubled, the accuracy increased by almost 10% and added 10 layers with 8 estimators in each layer. Summary of the investigation is provided in Table 1.

Table 1. Summary of the investigation conducted on the Deep Forest model with increase in number of classifiers

Database	Classifiers	Layers	Estimators	Accuracy
CK+	4	5	7	93.22%
CK+	8	10	7	93.22%
BU-3DFE	4	8	8	57.98%
BU-3DFE	8	10	8	65.53%

Table 2. The result comparison of FERAtt (Facial Expression Recognition with Attention Net) with Deep Forest learning

Author	Database	Method	Accuracy
Fernandez et al. [10]	BU-3DFE	FERAtt	75.22%
Our	BU-3DFE	Deep Forest	65.53%
Fernandez et al. [10]	CK+	FERAtt	86.67%
Our	CK+	Deep Forest	93.22%

5.1 Result

Figures 4 and 6 are the confusion matrices of the model probabilistic predictions accuracy on the BU-3DFE and CK+ respectively. Also, Figs. 5 and 7 are the graph of average recognition rate on the test data of BU-3DFE and CK+.

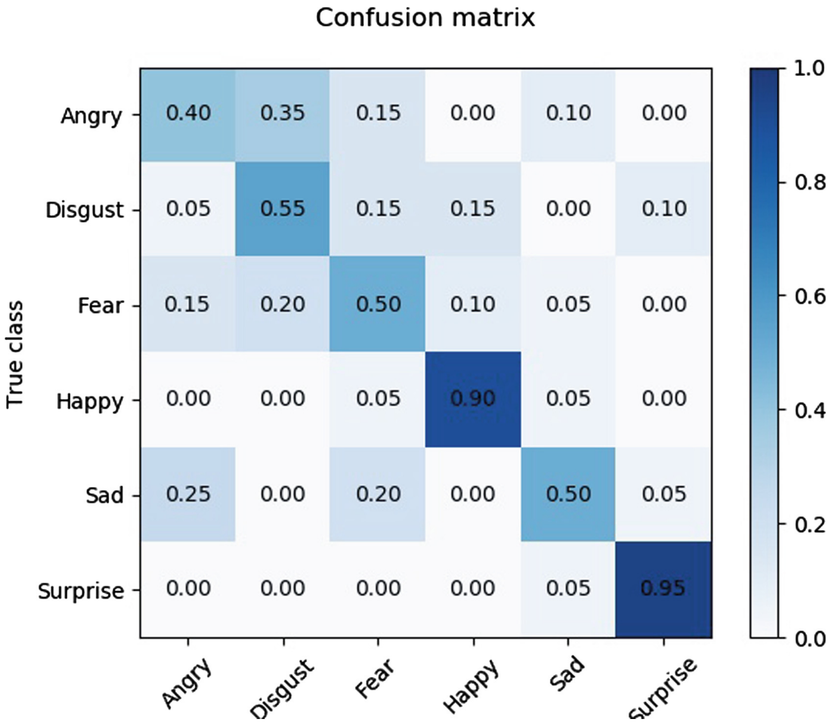


Fig. 4. Confusion matrix of Deep Forest predictions on BU-3DFE dataset

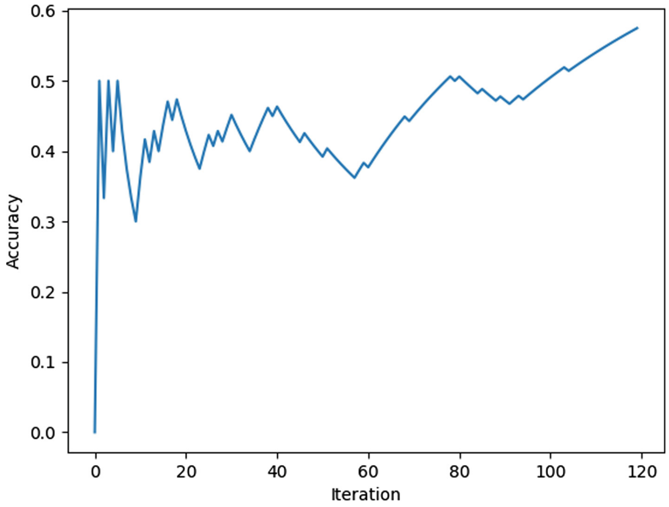


Fig. 5. The graph of the recognition rate against number of predictions of BU-3DFE test data

In Fig. 4, the prediction of the model is most for the surprise at 95%. Followed by happy at 90% then disgust at 55%, both sad and fear have 50% prediction accuracy and angry has the least prediction at 40%. Figure 6 shows that the model gives 100% prediction for Angry, disgust, Fear and happy instances, 94% for surprise and 40% for sad.

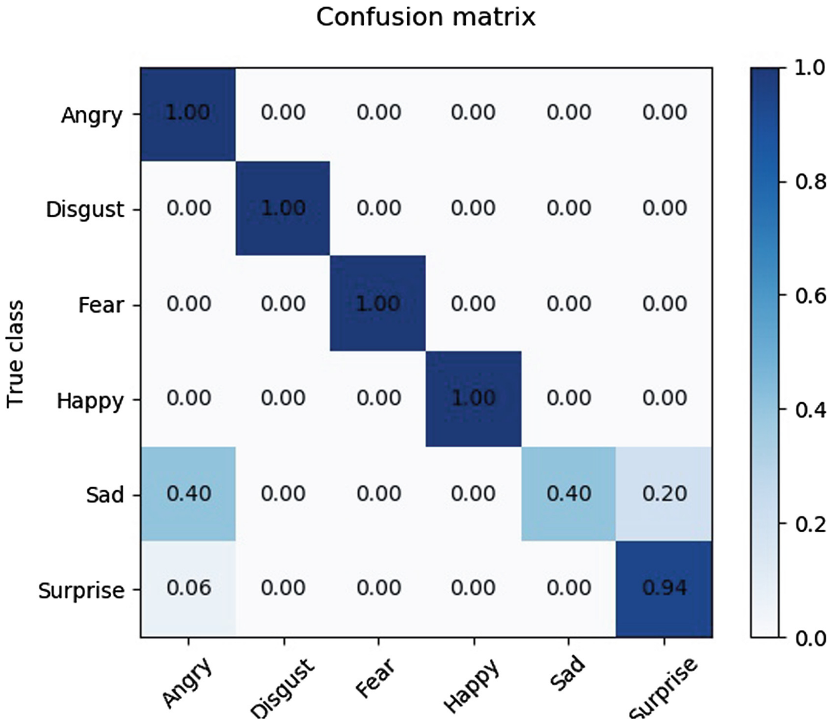


Fig. 6. Confusion matrix of Deep Forest predictions on CK+ dataset

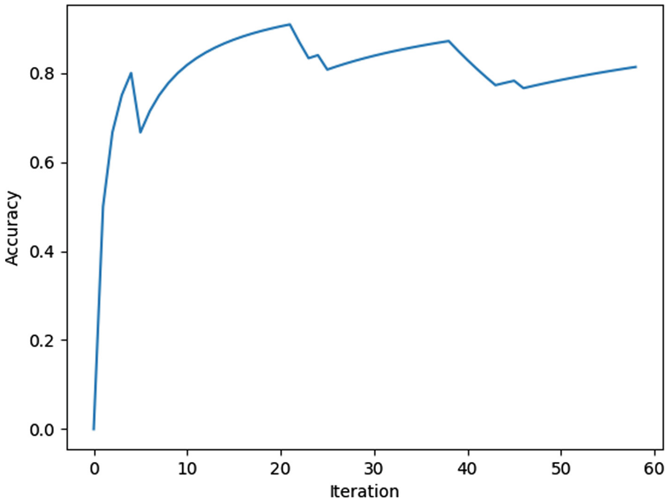


Fig. 7. The graph of the recognition rate against number of predictions of CK+ test data

We justify the performance of Deep forest on Facial expression classification by comparing its performance with the state of the art DNN method (FERAtt) [10]. Table 2, presents both our result and FERAtt result and clearly Deep forest gives better accuracy (93.22%) than the accuracy achieved in FERAtt (86.67%) on CK+ dataset. while accuracy gotten with FERAatt (75.22%) on BU3DFE dataset is more than Deep Forest (65.53%). But it should be noted that FERAtt could not use a small dataset, because the authors reported that the data were augmented and also combined with Coco data. Also FERAtt demands for high computing device like GPU for its appreciable time of computation, unlike the Deep Forest that performed its layer by layer learning on the available computing device (intel(R)Core(TM)i7-4770sCPU @3.10 GHz 3.10 GHz and RAM: 8 GB) at an appreciable time.

Obviously, the result of the experiment compliments the claim of [21]. It shows that Deep Forest has the inherent capability for small datasets. The average prediction accuracy of the model on CK+ (309 data) is 93.22% and BU-3DFE (600) is 65.53%. Although, Deep Forest is challenge with the issue of memory consumption, yet it could be a an alternative to DNN if its features are greatly explored.

6 Conclusion

We have presented a Deep learning approach other than the popularly known DNN for Facial Expression Recognition. And our work proved that Deep forest could preform very well even in a wild environment and with a sparsely distributed and unbalanced dataset. Also the outcome of the further investigation conducted in the experiment, is the evidence of dynamic control behaviour of deep forest over its model. The result of this work is an incite for exploring possibilities of enhancing Deep Forest model, which is the focus of the future work.

References

1. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Comput. **9**(7), 1545–1588 (1997). https://doi.org/10.1162/neco.1997.9.7.1545
2. Anggraeni, D., Wulandari, A., Barasaruddin, T., Yanti, D.: Enhancing CNN with preprocessing stage in automatic emotion enhancing CNN with preprocessing stage in automatic emotion recognition recognition. Procedia Comput. Sci. **116**, 523–529 (2017). https://doi.org/10.1016/j.procs.2017.10.038
3. Breiman, L.: Random forests. Mach. Learn. **45**(5), 1–33 (2001). https://doi.org/10.1023/A:1010933404324
4. Chen, J., Zhang, M., Xue, X., Xu, R., Zhang, K.: An action unit based hierarchical random forest model to facial expression recognition. In: International Conference on Pattern Recognition Application and Methods, ICPRAM, pp. 753–760 (2017). https://doi.org/10.5220/0006274707530760
5. Chollet, F.: Deep Learning with Python. Manning Publications Co., Shelter Island (2018)
6. Dapogny, A., Bailly, K., Dubuisson, S.: Pairwise conditional random forests for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 3783–3791 (2015). https://doi.org/10.1109/ICCV.2015.431

7. Deng, H., Zhu, J., Lyu, M.R., King, I.: Two-stage multi-class AdaBoost for facial expression recognition. In: IEEE International Conference on Neural Networks - Conference Proceedings, no. 1, pp. 3005–3010 (2007). https://doi.org/10.1109/IJCNN.2007.4371439
8. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45014-9_1
9. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**, 169 (1992). https://doi.org/10.1080/02699939208411068. http://www.tandfonline.com/loi/pcem20, http://dx.doi.org/10.1080/02699939208411068
10. Fernandez, P.D.M., Peña, F.A.G., Ren, T.I., Cunha, A.: FERAtt: facial expression recognition with attention net. In: IEEE Xplore (2019). http://arxiv.org/abs/1902.03284
11. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis the robotics institute. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000). https://doi.org/10.1109/AFGR.2000.840611
12. Kontschieder, P., Fiterau, M., Criminisi, A., Bul, S.R., Kessler, F.B.: Deep neural decision forests. In: International Joint Conference on Artificial Intelligence, pp. 4190–4194 (2016)
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, pp. 94–101, July 2010. https://doi.org/10.1109/CVPRW.2010.5543262
14. Miller, K., Hettinger, C., Humpherys, J., Jarvis, T., Kartchner, D.: Forward thinking : building deep random forests. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 1–8. NIPS (2017)
15. Vasanth, P.C., Nataraj, K.R.: Facial expression recognition using SVM classifier. Indones. J. Electr. Eng. Inf. (IJEEI) **3**(1), 16–20 (2015). https://doi.org/10.11591/ijeei.v3i1.126
16. Rokach, L., Maimon, O.: Classification trees. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 149–174. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_9. Chap. 9
17. Rosato, M.J.: A 3D facial expression database for facial behavior research a 3D facial expression database for facial behavior research. In: International Conference on Automatic Face and Gesture Recognition (FGR 2006), pp. 211–216, May 2006 (2016). https://doi.org/10.1109/FGR.2006.6
18. Salmam, F.Z., Madani, A., Kissi, M.: Facial expression recognition using decision trees. In: Proceedings - Computer Graphics, Imaging and Visualization: New Techniques and Trends, CGiV 2016, pp. 125–130. IEEE (2016). https://doi.org/10.1109/CGiV.2016.33
19. Su, M.C., Hsieh, Y., Huang, D.Y.: A simple approach to facial expression recognition. In: International Conference on Computer Engineering and Applications, pp. 458–461, January 2007 (2014)
20. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)
21. Zhou, Z.H., Feng, J.: Deep forest : towards an alternative to deep neural networks. In: International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 3553–3559 (2017)

3.2.5 Conclusion

Deep forests achieved a reduction in computation complexities with a promising recognition rate on the small volume of FER data (BU-3DFE and CK+). CK+ has a recognition rate above 90%, and the recognition rate of BU-3DFE is less than 70%. It could be deduced that deep forest performance degrades as the volume of data increases and with more challenging datasets.

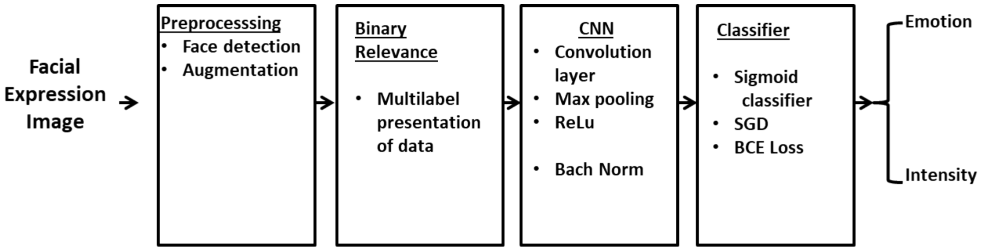


Figure 3.2: Block diagram description of the processes in ML-CNN.

3.3 Facial Expression recognition and ordinal Intensity Estimation: A Multilabel Approach

3.3.1 Introduction

The previous section presented a multitask implementation of FER. This section presents a deep multilabel framework that performs concurrent emotion recognition and the associate intensity estimation using ordinal metrics.

3.3.2 Method Discussion

Figure 3.2 presents the framework for a multilabel convolution network for facial expression recognition.

Preprocessing The data preparation method adopted in this work includes face localisation and data augmentation. With face localisation, only the region of interest in the face was captured, which helps to minimise redundant information from facial expression data. A modified Viola and Jones face detection algorithm with integral graph and AdaBoost algorithm was used to detect faces from the facial expression images. Augmentation is employed next to facial localisation to increase the data size and control data imbalance, especially in the CK+ dataset. Both facial localisation and data augmentation are implemented offline.

Binary Relevance Binary relevance is used to define the facial expression task as a multilabel problem. With Binary relevance, expression data are presented such that each expression image is associated with a degree of intensity.

Here, facial expression and intensity estimation tasks are formally presented as a multilabel problem. Generally, assume $X = R^m$ represents set of training samples with m dimensional feature vectors, a sample $x \in X$ associated with a label $y \in Y$ is given as $E = \{x_i; y_i\}$ such that $y_i \subseteq k$ where $k = \{y_i : i = 1.....p\}$ is the set of p possible labels. This implies that multilabel classification assigns more than one emotion to an expression image.

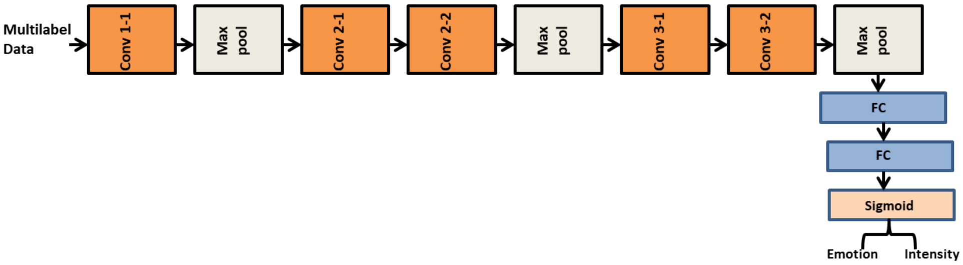


Figure 3.3: Order of operation of Convolution, Max Pooling, Fully connection and Sigmoid classifier in ML-CNN model.

In the context of facial expression recognition and intensity estimation, a special multilabel scenario is defined. An expression image is associated strictly with emotion information $y_i \in Y$ and intensity information $z_i \in Z$. Formally, given an Expression image E $= \{x_i, (y_i, z_i)\}$ where $y_i, z_i \in Y \times Z$. The challenge in this multilabel task is to generate a supervised classifier C capable of taking an unseen expression image E and simultaneously predicting its correct emotion state and intensity. IF $E = (x_i,)$ Then $C(E) \rightarrow Y \times Z$, which is the accurate emotion and intensity associated with the image. This transformation is achieved with binary relevance extension transformation technique as proposed by[35] with consideration for labels dependency. Binary relevance also aids in adopting deep learning into the multilabel environment.

Convolution Neural Network (CNN) CNN with a sigmoid output layer is the classifier, classifying facial expression images into emotion and the corresponding intensity as shown in Figure 3.3.

Convolution Operation: with convolution operation, patches from expression images are extracted and transformed to generate a 3D feature map where the depth is the number of filters that encode the unique part of the expression image. The convolution operation can learn local patterns from the image when the image is convoluted with the kernel. This model uses the kernel of size 3×3 with default stride and the output feature map specify at each convolution layer. Padding (zero paddings) is introduced to minimise border effects experienced during convolution. Equation (3.2) is the mathematical representation of the convolution operation whereby the input expression image is I(x,y), convoluted with kernel H (f,f) and the output feature map G[p,q]of p row and emphq is expressed accordingly. The convolution layer also employs an activation function, which is continuous and differentiable for learning a non-linear transformation of the input data and enhances the network to access a rich hypothesis space from deep representation. This work employs 3×3 kernel, ReLu activation function, zero-padding one stride and batch normalisation at each convolution layer. There are five convolution blocks in this model, and the first convolution layer convolutes the input image with the kernel to produce 32 feature maps, a non-linear activation function ReLu is applied to learn the non-linearity features, sparsity control and also to prevent gradient vanishing, which is likely to occur during back-propagation. For the stability of each layer, we also used batch normalisation and 0.5 dropouts. All these operations took place at each convolution layer, except the different filters generated at other convolution layers. At the second and third convolution layers, 64 features maps are produced, and at the fourth and fifth layers, 128 features maps are produced.

$$G[p,q] = (I * H)[x,y] = \sum_{j} \sum_{k} H[f,f] I[x-f, y-f] \tag{3.2}$$

Where I represents the facial expression image, H is the size of the filter x,y are the rows, and the column of the image feature, f,f are the rows, and the column of the filter, G[*x,y*] is the result of the convolution computation (Feature map).

Furthermore, the dimension D(p,q) of the output feature map as given in (3.3) is computed using the expression image size (m,n,c), s is strides number and filter H(*f,f*), m is the height of the image, n is the width, c is the channel or depth of the image and p is the number of padding.

$$D(P,Q) = [m,n,c] * [f,f,c] = \frac{m+2p-f}{s} + 1, \frac{n-2p-f}{s} + 1 \tag{3.3}$$

Pooling layer: this is a sub-sampling layer of the network where the down-sampling operation takes place. Its goal is to reduce feature maps' dimensions and ensure the preservation of representative features. Pooling operation reduces the computation complexity by reducing the number of training parameters, distortion, rotation, translation, and scaling sensitivity in the input. This system employs max-pooling methods. The max-pooling feature maps are convoluted with a 2 × 2 kernel to return the maximum value from each region covered by the kernel. Max-pooling operation is performed on the output of the convolution layer. This network contains three pooling layers, and the first pooling layer is positioned after the first convolution layer, the second and the third pooling layers are after the third and the fifth convolution layer respectively, as shown in Figure 3.3

Fully Connected layer: This layer behaves like a feed-forward network. The output of the last pooling layer is flattened; that is, the 2-dimensional matrix is unrolled into a vector. This is because a fully connected layer takes a one-dimensional matrix as input, the flattened function converts the height (h), the width (w) and the feature maps(f) into a series of feature vectors ($h \times w \times f$). The fully connected layer function is formally defined in (3.4) where x is the feature vector, $\gamma(.)$ is the ReLu activation function, and b is the network bias.

$$Fc_{w,b}(x) = \gamma(w^T x + b) \tag{3.4}$$

sigmoid function $S(.)$: The output layer of the model employed a sigmoid classifier, an activation function that can generate an independent probability for each of the classes, making it suitable for the multilabel classification task. $\mathrm{s_i}$ in

(3.5) represents the output of the fully connected layer. Stochastic gradient descent (SGD) is used as the model optimizer. While Binary cross-entropy in (3.6) compute the model loss. The order of convolution operation, pooling operation, the Fully connected and sigmoid operation is expressed in Figure 3.3

$$S(s_i) = \frac{1}{1+e^{-s_i}} \tag{3.5}$$

$$BCE(y,\overline{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y \cdot log(\overline{y_i}) - (1-y) \cdot log(1-\overline{y_i})) \tag{3.6}$$

in (3.5) y is the actual class label while the $\overline{y}_{\mathrm{i}}$ is the predicted label from the model.

3.3.3 Model Evaluation

The multilabel metrics for evaluating the ML-CNN model include; hamming loss, ranking loss, ranking average precision, and coverage loss.

Hamming loss H computes the loss generated between the binary string of the actual label and the binary string of the predicted label with XOR operation for every instance of the test data. The average overall sample is taken as given in (3.6) below.

$$H = \frac{1}{|N|.|L|}\sum_{i=1}^{|N|}\sum_{j=1}^{|L|} XOR(y_{i,j}, \hat{y_{i,j}}) \tag{3.7}$$

where $y_{\mathrm{i,j}}$ and $\hat{y}_{\mathrm{i,j}}$ are the ground truth and the predicted classes respectively.

Ranking loss captures pairs of emotion and intensity that are incorrectly ordered. The equation is presented in (3.7) Given that $y \in \{0,1\}^{\mathrm{N} \times \mathrm{k}}$ and the prediction of each label is described as: $\hat{}f \in \mathrm{R}^{(\mathrm{N} \times \mathrm{k})}$

$$Rank_{loss}(y, \hat{f}) = \frac{1}{N}\sum_{1=0}^{N-1}\frac{1}{||y_i||_0 k - ||y_i||_0}|Z| \tag{3.8}$$

k is the number of labels, N is the number of sample, and Z is (m,n): $\hat{}f_{\mathrm{i,m}} \geq \hat{\mathrm{f}}_{\mathrm{i,n}}$, $\mathrm{y}_{\mathrm{i,m}} = 1$, $\mathrm{y}_{\mathrm{i,n}}=0$, |.| is the cardinality of the set

Average precision: this metric is employed to find the higher-ranked function that is true for each ground-truth label. The higher the value within the closed range between 0 and 1, the better the model's performance. A mathematical illustration of label ranking average precision is given in (3.8) below. Given that $y \in \{0,1\}^{(\mathrm{N} \times \mathrm{k})}$ and the prediction of each label is described as: $\hat{\mathrm{f}} \in \mathrm{R}^{(\mathrm{N} \times \mathrm{k})}$ k is the number of labels, N is the number of sample,

$$LRAP = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{||y||_0}\sum_{j:y_{i,j}=1}\frac{|L_{i,j}|}{Ri,j} \tag{3.9}$$

where $L_{\mathrm{i,j}}$ = k: $\mathrm{y_{i,k}}$=1,$\mathrm{\hat{f}_{i,k}} \geq \mathrm{\hat{f}_{i,j}}$ and |.| is the cardinality of the set, $|.|_0$ computes the number of non zero element in a vector.

Coverage error: this metric evaluates, on average, the number of labels that should be included so that all the correct labels would be predicted at the final prediction. The smaller the value, the better the model performance. The mathematical illustration is shown in (3.9).

Given that $y \in \{0,1\}^{\mathrm{N} \times \mathrm{k}}$ and the prediction of each label is described as: $\mathrm{\hat{f}} \in \mathrm{R}^{\mathrm{N} \times \mathrm{k}}$

k is the number of labels, and N is the number of samples.

$$Coverage(y,\hat{f}) = \frac{1}{N}\sum_{i=0}^{N-1} maxrank_{i,j} \tag{3.10}$$

where $\mathrm{rank_{i,j}}$ is $|\{\mathrm{k}:\mathrm{\hat{f}_{i,k}} \geq \mathrm{\hat{f}_{i,k}}|\}$ and $\mathrm{j{:}y_{i,j}} = 1$

Result Presentation and Model Comparison ML-CNN performance is evaluated based on the multilabel metrics discussed. The results of each metric on the BU-3DFE and the CK+ datasets are detailed in the next section and in Chapter 4. The Model is also compared with some standard multilabel algorithms (RAKELD, CC, MLkNN and MLARAM). These algorithms are selected as bases for ML-CNN comparison because they are standard multilabel models in which the comparison could base on multilabel metrics. They are widely used for multilabel tasks and model comparison and are publicly available. Moreover, they were implemented on the same platform and evaluated using the same datasets to ensure fairness in comparison.

This paper has been presented in 15th International Symposium, ISVC 2020 San Diego, CA, USA, October 5–7, 2020 [1] and published in Ekundayo O., Viriri S. (2020) Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach. In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2020. Lecture Notes in Computer Science, vol 12510. Springer, Cham. https://doi.org/10.1007/978-3-030-64559-5_46 [2]

[1] 15th International Symposium, ISVC 2020 San Diego, CA, USA, October 5–7, 2020

[2] Ekundayo O., Viriri S. (2020) Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach. In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2020. Lecture Notes in Computer Science, vol 12510. Springer, Cham. https://doi.org/10.1007/978-3-030-64559-5_46

Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach

Olufisayo Ekundayo and Serestina Viriri(✉)

School of Mathematics, Statistics and Computer Sciences,
University of KwaZulu-Natal, Durban, South Africa
218085734@stu.ukza.ac.za, viriris@ukzn.ac.za

Abstract. Facial Expression Recognition has gained considerable attention in the field of affective computing, but only a few works considered the intensity of emotion embedded in the expression. Even the available studies on expression intensity estimation successfully assigned a nominal/regression value or classified emotion in a range of intervals. The approaches from multiclass and its extensions do not conform to man heuristic manner of recognising emotion with the respective intensity. This work is presenting a Multi-label CNN-based model which could simultaneously recognise emotion and also provide ordinal metrics as the intensity of the emotion. In the experiments conducted on BU-3DFE and Cohn Kanade (CK+) datasets, we check how well our model could adapt and generalise. Our model gives promising results with multilabel evaluation metrics and generalise well when trained on BU-3DFE and evaluated on CK+.

Keywords: FER · Multilabel · Ordinal · Intensity estimation

1 Introduction

Human face contains much information through which estimated parameters like; identity, age, emotion, gender, status, race and so on about an individual could be deduced. Facial expression as one of the non-verbal communication channels contains the most substantial proportion of man's medium of communication [6,15]. The main goal of the facial expression recognition system is to automate the inherent ability in human beings and detects the man affective state directly from changes experienced in the face, as a result of the facial muscular response to the affective state. The process of achieving this classification is known as Facial Expression Recognition (FER). Ekman and Friesen [3] studies aided facial expression classification into their proposed emotion states. Several techniques of achieving FER have been introduced in the literature. The list is not limited to handcrafted methods, conventional machine learning methods, Neural Network and deep learning methods. Studies in affective computing

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 581–592, 2020.
https://doi.org/10.1007/978-3-030-64559-5_46

considered expression recognition from a facial image in both static (controlled and uncontrolled) environment and dynamic environment. Automation of facial expression is vital in Human-Computer Interaction (HCI) and Computer Vision (CV). Areas of facial expression application keep evolving, and virtually applicable to every area where communication or human interaction with a system is involved.

Facial expression is subtle, [17,19,20] claimed that expression is often reflected as mixture of basic emotion in face. [25] emphasised that in the real world, the display of pure emotion is rare, and that emotion as a subjective notion should be assigned a relative value and not the absolute value of the standard classification algorithms. There is virtually no pure emotion because emotion is always accompanying by some other information that portrays its semantics. Limiting the recognition or detection of emotion subjectively to six basic emotion states (anger, disgust, fear, happy, sad and surprise) undermines the performance of facial expression recognition system.

One of the information that accompanies emotion is the degree or intensity of the expression displayed. It is undeniable that man recognizes emotion along with some ordinal metrics that depict the degree or the rate of expressing emotion in the face. To adequately capture the semantics of emotion, it is better to consider the ordinal information associated with it. Some of the research conducted on intensity estimation of facial expression images include [1,18]. Most of the existing approaches consider the task as a regression problem [22], which is far from man's concept of estimating emotion [15,25]. Man has a hierarchical structure perception about emotion and therefore estimate it using referenced base value, which allows its semantics preservation.

A man could identify the affective state with its accompanied relative intensity value from the face. Therefore, emotion intensity could appropriately be represented using ordinal metrics to the best of our knowledge non of the existing studies on emotion intensity estimation as regard facial expression considered ordinal metrics in their works. It is understandably, the lack of hierarchical or ordinal annotated data for facial expression intensity estimation task could be responsible for such limitation. Nevertheless, facial expression recognition should not be restricted to a multi-class problem. This argument is substantiated with the fact that more information about affect states are inferred from facial expression than the fundamental categorical values. This work considers facial expression recognition as a multi-label task, with the motive that an image of facial expression belongs to one of the six emotion states with an associated degree of intensity. The significance of this work is attributed to a relative based approach for emotion intensity estimation by using ordinal metrics. Another Uniqueness is the multi-label modelling and transformation of emotion recognition and intensity estimation, which has not been adopted for intensity estimation as far as our knowledge is concerned. This work is adapting CNN network to FER multilabel classification task, where both emotion and emotion intensity are simultaneously recognized.

The order of arrangement of this work is as follow; Sect. 2 is the review of some of the existing works on intensity estimation of facial expression recognition. It shall include a thorough elucidation of the methods and their respective limitations. Section 3 gives the general discussion of the deep learning classification model we are considering. Section 4 contains the description of the multilabel approach and the adaptation of CNN model to multilabel problem. Section 5 presented the description of the experiment and provides information of both the databases and the multilabel metrics for evaluation. Section 5.4 contains the result presentation and the discussion. Section 6 provides the conclusion of the work and the possible future works.

2 Related Works

Most of the research carried out on facial expression recognition, approach it from the perspective of multiclass problem [8,21]. [5,11,14] are recent review of different deep learning approaches on facial expression recognition, for any interested reader. Many successes in this field have been recorded with different machine learning algorithms and image processing techniques, especially with the state of the art method [4,16]. Regardless of the success, we cannot ignore the fact that facial expression image in most cases consists of more than one information about the affective state it is representing [20,25]. One of the extensions of facial expression recognition task is intensity estimation of facial expression recognition, and most of the studies in this regard are an extension of a multiclass or regression problem.

The popularly adopted techniques to facial expression intensity estimation are grouped into regression-based, distance-based, graphical-based and clustering-based model, as discussed in [9]. For instance Rudovic et al. [22] employed a manifold to modeled the topology of multidimensional continuous facial affect data by using a Supervised Locality Preserving Projection (SLPP) algorithm to encode the ordering of the expression class labels, to achieve smooth transitions between emotion intensities on the manifold. The topology was later incorporated into the Hidden Conditional Ordinal Random Field (H-CORF), and to ensure that the proposed dynamic ordinal regression is preserved, H-CORF parameters were constrained to lie on the ordinal manifold by forcing latent variables as a Gaussian Markov random field. The resulting model simultaneously achieved both dynamic recognition and intensity estimation of facial expressions of multiple emotions. Kimura and Yachida [23] proposed that facial expression recognition and degree estimation could be achieved through expressionless face referencing. They model the expressionless face with an elastic net model having the notion of obtaining any slight deformation caused by expression displayed in a face, with a motion vector of the deformed net. The motion vector of the node is mapped to low dimensional Eigenspace using K-L expansion, and estimation derived by the projection of the input image into the emotion space. Kamarol et al. [9] proposed a framework for facial expression recognition and intensity estimation with low computation. Feature extraction was carried-out using AAM (geometric feature), and the feature developed with KNN and weighting scheme; also the input video was represented by a weight vector that contained the most likely expression class of the input sequence and the expression intensity present in each frame in the sequence. HMM, as a classifier detects the emotion, and a change point detector encodes the expression intensity from the weight vector. The above mentioned compute intensity estimation using absolute value (quantitatively), this is not appropriate according to [25]. The goal of this work is to estimate facial expression intensity base on ordinal value and the emotion recognition simultaneously.

3 Convolution Neural Network

Convolution Neural Network is a deep learning model whose concept evolved from the Artificial Neural Network. It was first introduced by Lecun et al. [7]. CNN is an algorithm purposely designed for image processing and Computer vision tasks. Just like most deep learning networks, CNN performs an end to end learning, and the procedure executes in a hierarchy of layers. Each layer of CNN produces representation features ranging from low-level features of the image to a more abstractive concept. The process at which CNN automatically learns its representation features emulates the vision mechanism of an animal. That is, the animal visual cortex inspires CNN architectural design. CNN models are self-sufficient in extracting their representation features, and there is no need for any pre-calculated methods for features extraction. Its high performance contributes immensely to its popularity.

CNN could achieve end-to-end learning with the aid of the back-propagation algorithm and guided by loss function that leads the networks to the optimum result. Depending on the nature of the problem, softmax loss is mostly used in a multi-class problem where the chance of a class is dependent on the chances of occurrence of others. Recently sigmoid function initially meant for binary classification is adapted to multi-label tasks capitalising on its capability to generate a probability score of each available class independently. An instance of multi-label CNN network is discussed in [12] where the network adapted to the learning of topology preserved ordinal relationship and age difference information for age estimation and prediction task. DBCNN (Deep Bimodal Convolution Neural Network) proposed by Li and Deng [10] for facial expression recognition gives a promising result in the recognition of mixture or compound emotion from expression displayed in the face.

In this work, we are considering a multilabel CNN based network capable of learning the emotion features and the intensity of expression features for the prediction of emotion with its respective degree of intensity in an ordinal manner.

4 Multi-label Approach

In a multi-label classification problem, an image may belong to two or more categories of classes. The multi-label properties stated is evident in facial expression classification task; an expression image carries information of both the type of the emotion and the degree at which it is expressed. In this work, we considered a model transformation technique for the implementation of the facial expression recognition multi-label task.

4.1 Problem Description

Data description, Let $X = R^m$ represent the input space, and let $Y = y_1, \dots\dots\dots y_c$ denote the complete sets of label where c is the number of possible label value, also $Z = z_1, \dots\dots, z_k$ is the complete set of the degree where $k \in K$ is the intensity of the label $y \in Y$. Then the possibility of sample $x \in X$ having c_{th} class Y_c with the associated k_{th} intensity estimation degree Z_k is expressed as a function; $F\colon X \rightarrow Y \times Z$. Figure 1 is the pictorial description of how emotion is mapped with the degree of intensity expression.

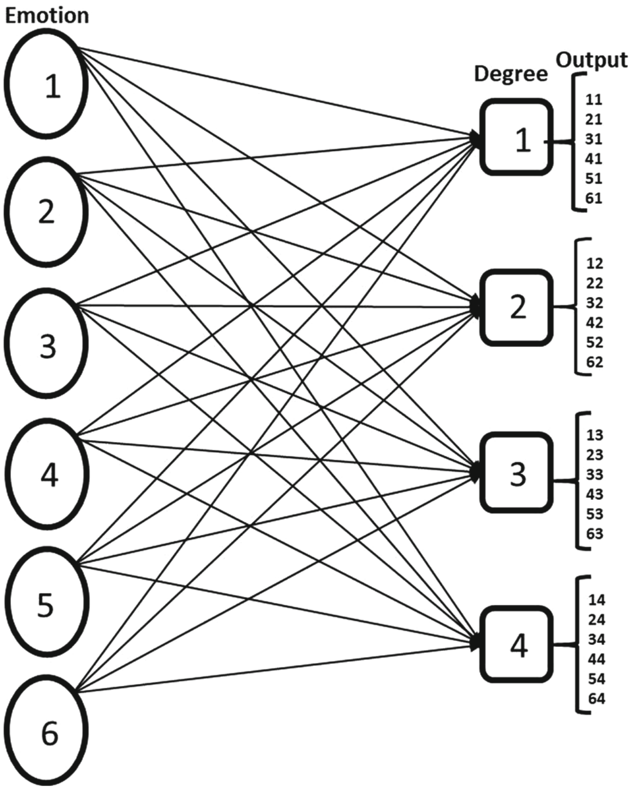


Fig. 1. The problem formulation of Multi-label Convolution Network showing the possible affective state with the respective degree of intensity of a facial expression sample. The nodes under emotion represent the six basic emotion classes Anger, Disgust, Fear, Happy, Sad, Surprise, the nodes under the degree Low, Normal, High, Very_High and the output is the possible result of the multi-label CNN classification.

This work employs a deep learning model to execute the multi-label task of facial expression recognition. The Convolution Neural Network architecture used is shown in the Fig. 2. The model is a variant of VGG-Network [24] in the arrangements of convolution layers. The model comprises of five convolution layers; the first convolution layer takes the input and convolutes it with a kernel of size 3 × 3 to give 32 filters as output. Max pooling with pooling 2 × 2 is applied for down-sampling after the first layer. Second convolution and third convolution layers output 64 filters each and likewise each of the forth and the fifth convolution layers output 128 filters. Down-sampling of the same pooling size is employed after the third and the fifth layer. The model has only one fully connected layer and the last layer, which is the sigmoid layer compute the likelihood of all the classes independently, which makes the FER multi-label implementation possible with CNN network. Some regularization techniques like drop-out and batch normalization are used at each layer in other to control the model from overfitting to the training samples.

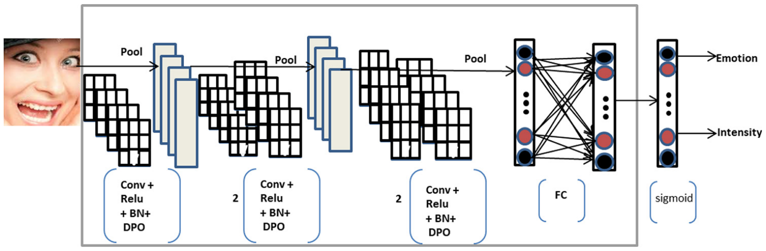


Fig. 2. The framework description of the proposed ML-CNN model: the network learns the emotion features and the degree of intensity features from the input image and make adjustments to the parameters with the aid of back-propagation during training. At the testing phase, the learned parameter of the network predicts the emotion and the respective intensity in the facial expression.

Convolution Operation: with convolution operation, patches from expression images are extracted and also transformed to generate a 3D feature map in which the depth is represented with the number of filters that encode the unique part of the expression image. The convolution operation can learn local patterns from the image when the image is convoluted with the kernel. This model uses the kernel of size 3 × 3 with default stride, and the output feature map is specified at each convolution layer. Padding (zero paddings) is introduced to minimize border effects experienced during convolution. Equation (1) is the mathematical representation of the convolution operation whereby the input expression image is f, convoluted with kernel h and the output feature map of x row and y column is computed.

$$G[x,y] = (f * h)[x,y] = \sum_{j}\sum_{k} h[j,k]f[x-j,y-k] \tag{1}$$

And the dimension of the output feature map (F) as given in (2) is computed using the expression image size (M, N), number of strides and filter (f, f).

$$D(p,q) = [m,n,c] * [f,f,c] = \frac{m+2p-f}{s} + 1, \frac{n-2p-f}{s} + 1 \tag{2}$$

Activation Operation: this module carries out the predictions base on the score generated from the feature map of each expression image. We use the ReLu activation function at each convolution layer of the network and the dense layer. We employed the sigmoid activation function in (3) for the final prediction at the fully connected layer and binary cross-entropy loss in (4) is used for the computation of the model loss.

$$S = \frac{1}{e^{-S}} \tag{3}$$

$$L(y, \overline{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y \cdot log(\overline{y_i}) - (1-y) \cdot log(1-\overline{y_i})) \tag{4}$$

in (4) y is the actual class label while the $\overline{y}$ is the predicted label from the model. Max pooling Operation: this operation is conducted on the output of the convolution layer. The feature map is downsampled when it is convoluted with the kernel size used is 2 × 2 and with two strides.

Optimization Operation and the regularization: Stochastic gradient descent (SGD) with a learning rate of 0.0001 is used as the model optimizer. Drop out and data augmentation are the employed regularization techniques for the model.

5 Experiment

This section is a brief description of the affective databases used, the preprocessing techniques, the experiment discussion and the multilabel evaluation metrics used.

5.1 Database

Binghamton University-3D facial Expression (BU-3DFE) dataset was introduced at Binghamton University by [26]; it contains 100 subjects with 2500 facial expression models. Fifty-six of the subjects were female, and 44 were male, the age group ranges from 18 years to 70 years old, with a variety of ethnic/racial ancestries, including White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino. 3D face scanner was used to capture seven expressions from each subject; in the process, four intensity levels were captured alongside for each of the six basic prototypical expressions. Associated with each expression shape model, is a corresponding facial texture image captured at two views (about +45° and −45°). As a result, the database consists of 2,500 two view's texture images and 2,500 geometric shape models. BU-3DFE dataset has been severally as a multi-label datasets [2,10] To the best of our knowledge BU-3DFE dataset is the only available dataset that annotated intensity of facial expression images in ordinal hierarchies. The intensity annotation of BU-3DFE is given in Fig. 3. Another popularly used benchmark dataset for intensity estimation is Cohn Kanade extension (CK+) data [13] This dataset, unlike the BU-3DFE, is a sequence data collected in a controlled environment. To make the CK+ conformable to our proposed method we adopted the general annotation of Onset, peak and offset in a frame to categorise the data into Low, Normal and High as the degree of emotional intensity.

5.2 Data Pre-processing and Experiment Discussion

This work considered two major and important pre-processing techniques to improve the performance of the system. We used facial landmarking algorithm to detect the region of the face and also to remove background information which could subject the system to unnecessary computation. To make the system robust against over-fitting because of the small quantities of available data for CNN to learn, we employed data augmentation. Data augmentation was not used on the fly to ensure data balancing especially in CK+ dataset.

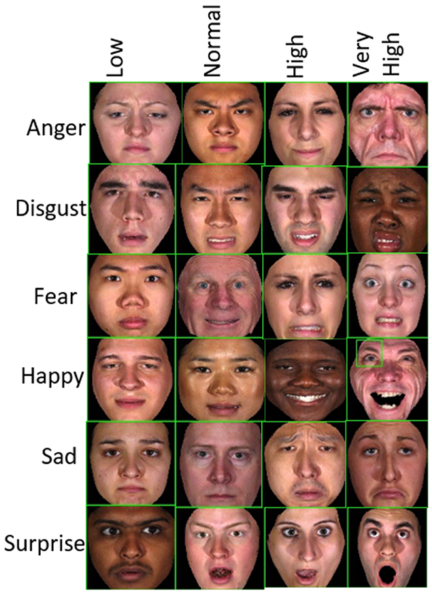


Fig. 3. This figure describes the hierarchical ordinal annotation of BU-3DFE, and provides information of six basic emotion states and their respective intensity estimation using relative value.

The experiment evaluated the proposed ML-CNN model on both BU-3DFE and CK+ datasets. After preprocessing, each of the raw data was scaled to a uniform size of 96 × 96. The datasets were first partitioned into the training set (70%), the validation set (20%) and the remaining 10% for the testing set. The training dataset was augmented and the pixel values were divided by 255 to ensure data scale normalization. We trained the proposed multi-label CNN network model on the training datasets. At each of the convolution layer of the network, we prevent over-fitting by using batch-normalization and dropout regularization techniques. We used Stochastic Gradient Descent with initial learning rate of 0.0003 for the network optimization. And Validation follows immediately using validation dataset. We made some investigations in the experiments. We observed the performance of our model on the raw data without augmentation, we likewise checked for the performance when face localization and augmentation were applied and lastly, we observed how well the model was able to generalise to unseen data samples by training the system with BU-3DFE and validate with CK+ data.

For each of our investigation, we conducted the model evaluation with the data testsets. The choice of our model performance metrics are the multilabel performance evaluation metrics; the hamming loss, coverage, ranking loss, and cross entropy loss. Brief discussion of the metrics are given in the next section.

5.3 Evaluation Metrics

Hamming loss computes the loss generated between the binary string of the actual label and the binary string of the predicted label with XOR operation for every instance of the test data. The average over all the sample is taking as given in (5) below

$$H = \frac{1}{|N|.|L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|J|} XOR(y_{i,j}, \hat{y_{i,j}}) \tag{5}$$

Ranking Loss: this metric is used to compute the average of numbers where the labels are incorrectly ordered. The smaller the ranking loss within the closed range between 0 and 1, the better our model performance. (6) is the mathematical illustration of ranking loss.

$$Rank_{loss}(y, \hat{f}) = \frac{1}{N} \sum_{1=0}^{N-1} ||y_i||_0 \frac{1}{k - ||y_i||_0} |Z| \tag{6}$$

where k is the number of labels and Z is (m, n): $\hat{\mathrm{f}}_{i,m} \geq \hat{\mathrm{f}}_{i,n}$, $\mathrm{y}_{i,m} = 1$, $\mathrm{y}_{i,n} = 0$.

Average Precision: this metric is employed to find the function of higher ranked that are true for each ground truth label. The higher the value within the closed range between 0 and 1 the better the performance of the model. Mathematical illustration of label ranking average precision is given in (7) below.

$$LRAP = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{||y||_0} \sum_{j:y_{i,j}=1} \frac{|L_{i,j}|}{Ri, j} \tag{7}$$

where $\mathrm{L}_{i,j}$ = K: $\mathrm{y}_{i,k} = 1$, $\hat{\mathrm{f}}_{i,k} \geq \hat{\mathrm{f}}_{i,j}$ and |.| is the cardinality of the set. Coverage error: this metrics is used to evaluate on average the number of labels that should be included so that all the true labels would be predicted at the final prediction. The smaller its value the better the model performance. The mathematical illustration is shown in (8).

$$Coverage(y, \hat{f}) = \frac{1}{N} \sum_{i=0}^{N-1} maxrank_{i,j} \tag{8}$$

where $\mathrm{rank}_{i,j}$ is —{k:$\hat{\mathrm{f}}_{i,k} \geq \mathrm{f}_{i,k}$—}.

5.4 Result and Discussion

We present a summary of the result obtained from four categories of the experiments in Table 1. The first experiment evaluates the proposed method on BU-3DFE without augmentation; the data contains 2400 samples. At the evaluation, we observed overfitting when the number of the epoch is 25, and as showed in Table 1. The coverage error 4.512 is high. The overfitting is traceable to insufficient data for the model to learn the representative features. In the other experiment that conducted augmentation on the training samples of BU-3DFE, no overfitting observed, and the values obtained for the metrics are promising. The result obtained from CK+ is relative to the BU-3DFE+Augmentation, which indicate the adaptability of the proposed method. The proposed method also generalised well when trained with BU-3DFE and tested with CK+, the value obtained for the hamming loss is 0.1668, the ranking loss is 0.1925, average precision is 0.7322, and the coverage error is 1.4810.

Table 1. The summary of the model performance evaluation using four multi-label metrics: ↑ indicates that the higher the value the better the model performance and ↓ indicates that the lower or smaller the value the better the performance.

Database	Hamming loss ↓	Ranking loss ↓	Average precision ↑	Coverage ↓
BU-3DFE	0.1521	0.4773	0.7353	4.512
BU-3DFE + AUG	0.0694	0.1911	0.8931	1.9457
CK+	0.1426	0.0581	0.8993	1.6473
Generalization	0.1668	0.1925	0.7322	1.4910

6 Conclusion

This work presented a novel approach which relied on the human unique means of detecting emotion and estimating the intensity simultaneously. Our proposed Ordinal multi-label deep learning based method gives a promising result on BU-3DFE- a multilabel benchmark dataset with intensity annotation in hierarchy of ordinal values. We extend the evaluation of our method also to a sequence dataset (CK+), the result was also a prospective one as indicated in Table 1. Our model was able to generalise well when it was trained with BU-3DFE and evaluated on CK+. The future work will tends towards model performance enhancement, and to consider the a dynamic/temporal and emotion in the wild environment using our reference based approach for emotion detection and intensity estimation.

References

1. Chang, K.Y., Chen, C.S., Hung, Y.P.: Intensity rank estimation of facial expressions based on a single image (2013). https://doi.org/10.1109/SMC.2013.538
2. Dhall, A., Goecke, R., Gedeon, T.: Automatic group happiness intensity analysis. IEEE Trans. Affect. Comput. **6**(1), 13–26 (2015). https://doi.org/10.1109/TAFFC.2015.2397456
3. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Social Psychol. **17**(2), 124–129 (1971). https://doi.org/10.1037/h0030377
4. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019). https://doi.org/10.1109/ACCESS.2019.2917266
5. Ghayoumi, M.: A quick review of deep learning in facial expression. J. Commun. Comput. **14**, 34–38 (2017). https://doi.org/10.17265/1548-7709/2017.01.004
6. Ghimire, D., Lee, J.: Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensor **13**, 7714–7734 (2013). https://doi.org/10.3390/s130607714
7. Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **551**, 541–551 (1989). https://doi.org/10.1162/neco.1989.1.4.541
8. Huang, Y., Chen, F., Lv, S., Wang, X.: Facial expression recognition: a survey. SS Symmetry **11**, 1189 (2019). https://doi.org/10.3390/sym11101189
9. Khairuni, S., Kamarol, A., Hisham, M., Kälviäinen, H., Parkkinen, J., Parthiban, R.: Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. Pattern Recogn. Lett. **92**, 25–32 (2017). https://doi.org/10.1016/j.patrec.2017.04.003
10. Li, S., Deng, W.: Blended emotion in-the-wild: multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. Int. J. Comput. Vis. **127**, 884–906 (2018). https://doi.org/10.1007/s11263-018-1131-1
11. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **3045**(c), 1–20 (2020). https://doi.org/10.1109/TAFFC.2020.2981446
12. Liu, H., Lu, J., Feng, J., Zhou, J.: Ordinal deep feature learning for facial age estimation. In: Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017–1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge, pp. 157–164 (2017). https://doi.org/10.1109/FG.2017.28
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, July, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
14. Mahmood, Z., Muhammad, N.: A review on state-of-the-art face recognition approaches. Fractals **25**(2), 1–19 (2017). https://doi.org/10.1142/S0218348X17500256
15. Martinez, B., Valstar, M.F.: Advances, challenges, and opportunities in automatic facial expression recognition. In: Kawulok, M., Celebi, M.E., Smolka, B. (eds.) Advances in Face Detection and Facial Image Analysis, pp. 63–100. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25958-1_4

16. Mayya, V., Pai, R.M., Pai, M.M.: Automatic facial expression recognition using DCNN. Proc. - Proc. Comput. Sci. **93**(September), 453–461 (2016). https://doi.org/10.1016/j.procs.2016.07.233
17. Xia, X.: Facial expression recognition based on monogenic binary coding. Appl. Mech. Mater. **512**, 437–440 (2014). https://doi.org/10.4028/www.scientific.net/AMM.511-512.437. Scientific Net
18. Nomiya, H., Sakaue, S., Hochin, T.: Recognition and intensity estimation of facial expression using ensemble classifiers. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1–6 (2016). https://doi.org/10.1109/ICIS.2016.7550861
19. Ekman, P., Friesen, W.V.: Unmasking the Face: A Guide to Recognising Emotion from Facial Clue. Malor Books, San Jose (2003)
20. Plutchik, R., Bering, J.M., Descriptions, B.: Contents a phylogenetic approach to religious origins on the subcortical sources of basic human emotions and emergence of a unified mind science, vol. 08191 (2001)
21. Ramzan, M., Khan, H.U., Awan, S.M., Ismail, A., Ilyas, M., Mahmood, A.: A survey on state-of-the-art drowsiness detection techniques. IEEE Access **7**, 61904–61919 (2019). https://doi.org/10.1109/ACCESS.2019.2914373
22. Rudovic, O., Pavlovic, V., Pantic, M.: Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2634–2641 (2012). https://doi.org/10.1109/CVPR.2012.6247983
23. Kimura, S., Yachida, M.: Facial expression recognition and its degree estimation, pp. 295–300 (1997). https://doi.org/10.1109/CVPR.1997.609338
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–14 (2015)
25. Yannakakis, G.N., Cowie, R., Busso, C.: The ordinal nature of emotions, pp. 248–255 (2017). https://doi.org/10.1109/ACII.2017.8273608
26. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 211–216 (2006). https://doi.org/10.1109/FGR.2006.6

3.3.4 Conclusion

Multilabel approach to FER in recognition of Facial expression and the corresponding intensity was achieved in this section, and the performance evaluation with some multilabel metrics reveal the proficiency of the proposed method. Nevertheless, there is still the possibility of achieving robustness of the method to overfitting and intraclass and interclass variations.

3.4 Multilabel Convolutional Neural Network for Facial Expression Recognition and ordinal intensity Estimation

3.4.1 Introduction

In this section, we present a multilabel framework for concurrent recognition of emotion and its intensity. Multilabel Convolution Neural Network (ML-CNN) employed binary relevance to model the emotion with the respective intensity using ordinal metrics. To minimise the overfitting effect, VGG-16, a pretrained Convolutional neural network, was employed. ML-CNN performance was also improved with island loss aggregation and binary cross-entropy loss, which minimised intraclass and maximised interclass variations.

Part of this paper has been presented [1] and published [2] and Enhanced method has been accepted for publiccation [3]

[1] 15th International Symposium, ISVC 2020 San Diego, CA, USA, October 5–7, 2020

[2] Ekundayo O., Viriri S. (2020) Facial Expression Recognition and Ordinal Intensity Estimation: A Multilabel Learning Approach. In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2020. Lecture Notes in Computer Science, vol 12510. Springer, Cham. https://doi.org/10.1007/978-3-030-64559-5_46

[3] Olufisayo Ekundayo and Serestina Viriri 2021, "Multilabel Convolutional Neural Network for Facial Expression Recognition and Ordinal Intensity Estimation", PeerJ computer science (under peer-review)

PeerJ Computer Science

Multilabel convolution neural network for facial expression recognition and ordinal intensity estimation

Olufisayo Ekundayo and Serestina Viriri

Computer Science Discipline, University of KwaZulu-Natal, Durban, South Africa

ABSTRACT

Facial Expression Recognition (FER) has gained considerable attention in affective computing due to its vast area of applications. Diverse approaches and methods have been considered for a robust FER in the field, but only a few works considered the intensity of emotion embedded in the expression. Even the available studies on expression intensity estimation successfully assigned a nominal/regression value or classified emotion in a range of intervals. Most of the available works on facial expression intensity estimation successfully present only the emotion intensity estimation. At the same time, others proposed methods that predict emotion and its intensity in different channels. These multiclass approaches and extensions do not conform to man heuristic manner of recognising emotion and its intensity estimation. This work presents a Multilabel Convolution Neural Network (ML-CNN)-based model, which could simultaneously recognise emotion and provide ordinal metrics as the intensity estimation of the emotion. The proposed ML-CNN is enhanced with the aggregation of Binary Cross-Entropy (BCE) loss and Island Loss (IL) functions to minimise intraclass and interclass variations. Also, ML-CNN model is pre-trained with Visual Geometric Group (VGG-16) to control overfitting. In the experiments conducted on Binghampton University 3D Facial Expression (BU-3DFE) and Cohn Kanade extension (CK+) datasets, we evaluate ML-CNN's performance based on accuracy and loss. We also carried out a comparative study of our model with some popularly used multilabel algorithms using standard multilabel metrics. ML-CNN model simultaneously predicts emotion and intensity estimation using ordinal metrics. The model also shows appreciable and superior performance over four standard multilabel algorithms: Chain Classifier (CC), distinct Random K label set (RAKEL), Multilabel K Nearest Neighbour (MLKNN) and Multilabel ARAM (MLARAM).

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Data Mining and Machine Learning

Keywords Binary cross-entropy, Facial expression recognition, Island loss, Multilabel, Ordinal intensity estimation

INTRODUCTION

Recognising human affective state from a facial image is one of the most relevant challenges in Computer Vision (CV) and Human-Computer Interaction (HCI). This aspect of Computer Vision has gained much attention; several methods and approaches have been proposed in the literature. Early methods resolved that FER is a multiclass

Submitted 8 April 2021
Accepted 13 September 2021
Published 29 November 2021

Corresponding author
Serestina Viriri, viriris@ukzn.ac.za

Academic editor
Alex James

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj-cs.736**

How to cite this article Ekundayo O, Viriri S. 2021. Multilabel convolution neural network for facial expression recognition and ordinal intensity estimation. **PeerJ Comput. Sci. 7:e736 DOI 10.7717/peerj-cs.736**

problem and thus proposed multiclass based classifiers or adapted binary classifier to multiclass problems as appropriate methods for FER classification. For instance, *Ekman & Friesen (1971)* categorised facial expression into six basic emotion classes: Anger, Disgust, Fear, Happy, Sadness and Surprise. This classification automatically restricted FER into a multiclass task and buried much information that could help achieve robustness and better accuracy. The concept of arousal and valence model reveals more information content of FER. While arousal considers the expression intensity, valence captures the pleasantness and the unpleasantness of the expression (*Mollahosseini, Hasani & Mahoor, 2019*; *Yang & Sun, 2017*).

The expression intensity can be classified as one of the main attributes of emotion in facial expression. *Plutchik (2001)* ascertained that expression is a result of combination of basic emotions in the face. *Yannakakis, Cowie & Busso (2017)* reiterated that in real life, the display of pure emotions is rare and described emotion as a relative notion that should not be classified in terms of absolute values in the standard classification algorithms. Expression recognition and intensity estimation is a common task executed by human beings. Human beings find it easy, convenient, and comfortable to predict the emotional state concurrently and the accompanying intensity (using ordinal metrics) of a person from expression image. This intrinsic ability in human has not been adequately modeled in FER system. The classification of facial expression into basic emotion states has been considered severally in diverse ways in the literature, yet the approach could not account for the intensity of the recognised emotion. Likewise, few studies on emotion recognition and intensity estimation from face image succeeded in assigning numeric values as the estimated intensity. This attempt is far from the perception of man towards emotion intensity estimation. Man has a hierarchical structure perception about emotion and therefore estimate it using referenced base value, which allows its semantics preservation. To the best of our knowledge, none of the works on facial expression recognition and intensity estimation considered static FER dataset, and the environments explored in the study are sequence and dynamic environments. The notion that sequence and dynamic data contain more information of expression intensity and lack of hierarchical annotated static dataset may be the cause. Our findings show that the only static dataset in the field with ordinal annotation is BU-3DFE.

In this study, FER is considered a multilabel problem because an instance of a facial expression image contains information about emotion displays and the corresponding intensity. The six possible emotion states include Anger, Disgust, Fear, Happy, Sadness and Surprise. The ordinal metrics for estimating the category of emotion intensity are: low, normal, high and v_high (very high). The first phase of the FER multilabel approach is data organisation. We organise the data such that each emotional state is associated with the corresponding intensity; this is pictorially represented in Fig. 1. We implement a problem transformation technique using binary relevance (BR). The CNN network with sigmoid function in the output layer serves as the binary classifier. Because of our dataset population, we use the pre-trained network (VGG-16) to avoid model overfitting, which a challenge in *Ekundayo & Viriris (2020)*. To reduce intraclass variation and increase interclass variation an aggregation loss (combination of island loss and BCE loss)

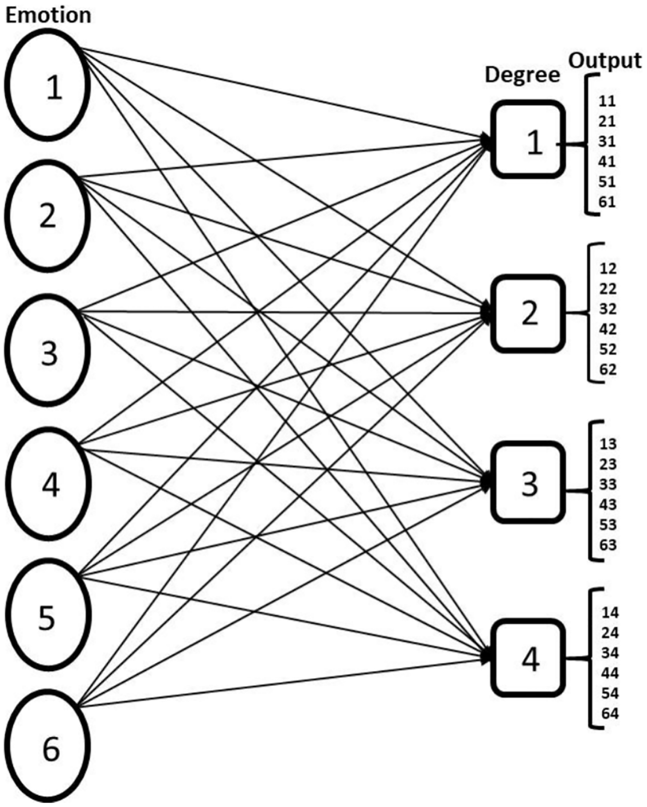


Figure 1 Showing multilabel problem formulation of FER. The nodes under emotion represent the six basic emotion classes Anger, Disgust, Fear, Happy, Sad, Surprise, and the nodes under the degree represent the ordinal estimation of emotion intensity Low, Normal, High, Very High and the output is the possible result of the multilabel CNN classification. Full-size DOI: 10.7717/peerj-cs.736/fig-1

is proposed which is another additional feature to our work in (*Ekundayo & Viriris, 2020*). The contributions of this work include:

- Multilabel model of facial expression recognition and intensity estimation. With this model, both the emotion features and the hierarchical structure embedded in them are learned concurrently.
- Ordinal metrics are used for the emotion intensity estimation, enabling the model to present the intensity estimation in a similar way like human beings.
- Use of Binary relevance multilabel transformation technique and CNN classifier, CNN is used as a binary classifier by implementing sigmoid function at the network's output layer. This ensures that the prediction probability of any class is independent of the other classes. Classifier sensitivity to intraclass and interclass variation is enhanced with the aggregation of island loss and BCE loss.

The proposed ML-CNN facial expression recognition model is capable of predicting the emotion and the corresponding ordinal intensity estimation concurrently from facial expression images. The simultaneous prediction of emotion and its intensity is a vital information in the application of FER; especially in psychiatry and schizophrenia

(*Behere, 2015*; *Seo et al., 2020*) and also for pain (*Chen, Ansari & Wilkie, 2012*; *Roy et al., 2016*) and depression analysis (*Guo et al., 2021*). Application of FER intensity estimation in real- world could mitigate the challenges of recognising emotion in schizophrenia patients, also since pain and depression are categorised as compound emotions (*Du & Martinez, 2015*) FER intensity estimation could appropriately state the degree to which they are expressed. Quantifying emotion with ordinal metrics makes ML-CNN to be similar to human prediction of emotion, which agrees with adaptation level theory account of Russell on emotion (*Russell & Lanius, 1984*), and ordinal nature of emotion as presented in *Yannakakis, Cowie & Busso (2017)*.

This work is organised as follows: Section "Related Works" discusses some studies related to both facial expression recognition and intensity estimation, the discussion also covers some of CNN network optimisation techniques. Section "Multilabel Convolution Neural Network Model (ML-CNN) Description" presents the ML-CNN model description; starting from problem formulation to describing the CNN network and the enhance loss functions employed. Section "Experiment" contains details of the experiments, which involve: the preprocessing of the data, and the experiment procedure details, and brief introduction of the databases. In Section "Experimental Results and Discussion", we provide a logical presentation of the experiments' result and relevant discussion of the experiments' outcomes. Section "Conclusions" is the conclusion of the work.

RELATED WORKS

In quest of a robust FER system, several studies have been conducted using traditionally handcrafted methods (*Li & Deng, 2020*; *Turan & Lam, 2018*), conventional machine learning techniques (*Ekundayo & Viriri, 2019*) and the state-of-the-art deep learning methods (*Liu et al., 2017*). The named techniques have been thoroughly considered under the supervised and unsupervised approaches in either a static or dynamic environment. Most of these approaches only succeeded in classifying an expression image into six or seven emotion classes.

Deep learning methods continue to evolve in diverse ways to achieve an optimal result in FER classification, and this is evident in EMOTIW2015, and EMOTIW2016 competition (*Fan et al., 2016*; *Kahou et al., 2015*). This section will concentrate more on the deep learning approach to facial expression recognition and intensity estimation, and some optimisation techniques adapted to CNN performance improvement.

FER classification is further extended to expression intensity estimation; few works on emotion intensity estimation are available in the literature; many works concentrate more on action unit intensity estimation. For example; *Gudi et al. (2015)* proposed a single CNN network for simultaneous estimation of Action Unit (AU) activation and intensity estimation. They claimed that activating the specific neuron of the output layer could result into a binary and continuous classification of AUs and corresponding intensity. Likewise, *Batista et al. (2017)* proposed AUMP Network (AUMP-NET), this network is a single network with multi-output regression capacity to learn AUs relationship and their respective intensity. The network is capable of learning the available

Table 1 **Summary of various models for emotion and intensity recognition.**

Method	Model	DB & performance	Limitation
Verma et al. (2005)	Distance based	Primary source: NA	Only few emotions are considered, method not generalise, emotion intensity before emotion recognition, computationally expensive.
Lee & Xu (2003)	Optical flow tracking algorithm (Distance)	Real-time data	Need for each subject to be trained differently, not generalise, predicting intensity before emotion
Kim & Pavlovic (2010)	HCORF (Prob)	CMU	Intrinsic topology of FER data is linearly model.
Quan, Qian & Ren (2014)	K-Means (Cluster)	CK+	Predict intensity before emotion, intensity estimation based on graphical difference is not logical
Chang, Chen & Hung (2013)	Scatering transform + SVM (Cluster)	CK+	Emotion recognition task is omitted.
Zhao et al. (2016)	SVOR (Regression)	Pain	Correlations between emotion classes are not modelled.
Rudovic, Pavlovic & Pantic (2012)	LSM-CORF (Prob)	BU-4DFE, CK+	Latent states are not considered in the modeling of sequences across and within the classes
Walecki et al. (2015)	VSL-CRF (Prob)	CK+ AFEW	Result of emotion intensity is not accounted for.
Kamarol et al. (2017)	weighted vote	CK+	Emotion and emotion intensity not concurrently predicted.
Proposed model	ML-CNN (Multi-Label)	BU-3DFE	Assume temporal information among sequence data as ordinal metrics.

Note:
NA: Not Applicable, MAE: Mean Absolute Error, PCC: Pearson Correlation Coefficient, ICC: Intraclass Correlation, MAL: MeanAbsolute Loss, HL: Hamming Loss, RL: Ranking Loss; AP: Average Precision, CE: Coverage Error.

AU and its corresponding intensity, simultaneously. Also, the network could learn to pose feature variations using multitask loss. These methods only determined the occurrence of AUs; the intensity is computed by regression means. The intensity of the AUs is not modelled in the training of the network. Similar studies on AU detection and intensity estimation could be found in Zhao et al. (2016) and Zhou, Pi & Shi (2017).

The few works on emotion recognition and intensity estimation are categorised in Kamarol et al. (2017) as: the distance-based (Verma et al., 2005), the cluster-based (Quan, Qian & Ren, 2014), the graphical-based (Valstar & Pantic, 2012) and the regression-based (Nomiya, Sakaue & Hochin, 2016) methods. As stated earlier, our focus is on recent deep learning approaches to emotion recognition and intensity estimation. Aamir et al. (2020) proposed a multilevel convolution neural network for expression classification and intensity estimation. The proposed deep network has two net phases: the expression-network phase, which handles the classification of facial expression image into the basic classes of emotion, and the intensity-network phase that takes the output of expression-network, which is one of the basic emotion and focus on the determination of the degree at which the recognised emotion is expressed. Summary of the existing method are presented in Table 1.

Xu et al. (2020) proposed a multitasking learning system using a cascaded CNN, and the objectives tend towards incorporating students attentiveness and students emotion recognition and intensity estimation into an intelligent class system. The first module of the cascaded network handled the preprocessing stages that involve face detection, face alignment and head pose estimation through which attentiveness is determined.

The second module implements an unsupervised raking CNN network to recognise the emotion and intensity estimation using ordinal evolution in the sequence data.

All of the stated approaches fail to adequately model the human mental capacity of predicting emotion with their respective intensity. The methods either estimate emotion intensity without emotion recognition or recognise emotion and its intensity separately. None of the methods carries out both tasks simultaneously. Multilabel learning is the recent trending approach to FER. This approach emerges from the public opinion that facial expression image contains a mixture of emotion, and only in a rare occasion is pure emotion displayed in face (*Plutchik, 2001*; *Yannakakis, Cowie & Busso, 2017*).

Facial expression challenges influenced FER system's performance, and the efforts in the field tend towards how the challenges could be reduced to bearable minimal. In the FER research community, diverse approaches have been implemented to enhance or optimise CNN networks to mitigate FER challenges. Some of the CNN optimisation approaches focus on improving the network's discriminating power through modification of loss function to reduce intraclass variance and increase interclass variance. Loss function guides the optimisation function in the direction to follow, and it states how close or far is the model prediction to the ground truth. The traditional loss function for multiclass tasks is softmax loss (*Liu et al., 2016*; *Wang et al., 2018*). The challenge identified with softmax loss is that while penalising the misclassified samples, it repels different classes to cluster apart, which is a challenge in FER, the introduction of center loss function aid to alleviate softmax loss challenge in the sense that it was able to cater for intraclass variation but fails to consider interclass variation appropriately. As discussed in *Cai et al. (2018)*, island loss is capable of increasing network discriminating power by increasing interclass variation and reducing intraclass variation, which is the main challenge in FER tasks. The experiment conducted in *Cai et al. (2018)* island loss function shows a better performance than either softmax loss or centered loss function. Likewise, *Li & Deng (2019)* in their effort to implement a robust FER with high discriminating power, form a tuplet cluster loss function, which is a hybrid of a tuplet (N+1) loss function and cluster loss function. The (N+M) tuplet cluster loss described an N-negative and M-positive sample in the CNN framework's minibatch. The formed tuplet cluster is combined with softmax loss as a joint optimisation technique to explore identity label and expression label information potentials thoroughly.

Other modification of CNN networks is found in *Alenazy & Alqahtani (2020)*, *Ozcan & Basturk (2020)*, *Wu, Wang & Wang (2019)* and *Zatarain Cabada et al. (2020)*. *Ozcan & Basturk (2020)* improved FER system performance with transfer learning and hyperparameter tuning. *Alenazy & Alqahtani (2020)* present a semi-supervised deep belief network for FER and employed gravitation search algorithm for network parameter optimisation. *Wu, Wang & Wang (2019)* optimise CNN network for FER classification by converting the output layer tensor of the network into a multidimensional matrix-vector *via* matrix transformation to enlarge the eigenvalues such that the system might have lower loss rate. *Zatarain Cabada et al. (2020)* proposed a genetic algorithm optimisation technique for CNN hyperparameter tuning for FER. The main goal of the genetic algorithm is to achieve the best solution from the hyperparameter population evolution.

This work is presenting an enhanced ML-CNN model for emotion recognition and ordinal intensity estimation. The proposed multilabel deep learning model can learn the hierarchical structure in FER datasets during the training of the network and predicts the emotion and the ordinal intensity in the expression face concurrently. Transfer learning optimisation technique is used as a trade-off for the insufficient data population for the appropriate ML-CNN model learning. The entropy loss function is fortified with island loss function to minimise the intraclass and interclass challenges. Detail description of our model is presented in the next section.

MULTILABEL CONVOLUTION NEURAL NETWORK MODEL (ML-CNN) DESCRIPTION

Deep learning models are traditionally employed in solving either a binary class or multi-class problems, where an instance of a population is only restricted to a group of class. In such a multitask challenge a single output is generated. Very few studies considered deep learning for multilabel tasks. *Liu et al. (2016)* practically established this fact that facial expression in the real world is more of mixture or compound of emotion. Their work verified this while using the Expectation-Maximization (EM) algorithm to automate the manual annotation of Real-world Affective Faces (RAF) database. Their approach shows that expression face contains more than one emotional state in different intensity level.

ML-CNN is a deep learning model we consider for classifying expression images into the emotional states and the associated degree of intensity. ML-CNN model combines multilabel problem transformation techniques with CNN algorithm as a deep learning technique for the multilabel classification task. Details of this model are considered in the following subsections.

Problem transformation

Here, we formally present facial expression and intensity estimation task as a multilabel problem. Generally, assume $X = R^m$ represents set of training samples with m dimensional feature vectors, a sample $x \in X$ associated with a label $y \in Y$ is given as $E = \{x_i; y_i\}$ such that $y_i \subseteq k$ where $k = \{y_i: j = ,\ldots p\}$ is the set of p possible labels. In the context of facial expression recognition and intensity estimation, a special multilabel scenario is defined. An expression image is associated strictly with emotion information $y_i \in Y$ and intensity information $z_i \in Z$. Formally, given an Expression image $E = \{x_i, y_i \times z_i\}$ where $y_i \times z_i \in Y \times Z$ such that $k_1 = \{y\}_{i=1}^{P}$ for all possible $p \in Y$ and $k_2 = \{z\}_{i=1}^{q}$ for all possible $q \in Z$. The challenge in this multilabel task is to generate a supervised classifier C which is capable of taken an unseen expression image E and simultaneously predict its correct emotion state and its intensity. That is, given $E = (x_i,)$ Then $C(E) \rightarrow Y \times Z$, which is the accurate emotion and intensity associated with the image. This transformation is achieved with binary relevance extension transformation technique as proposed by *Luaces et al. (2012)* with a slight modification that limits label independence. Binary relevance also

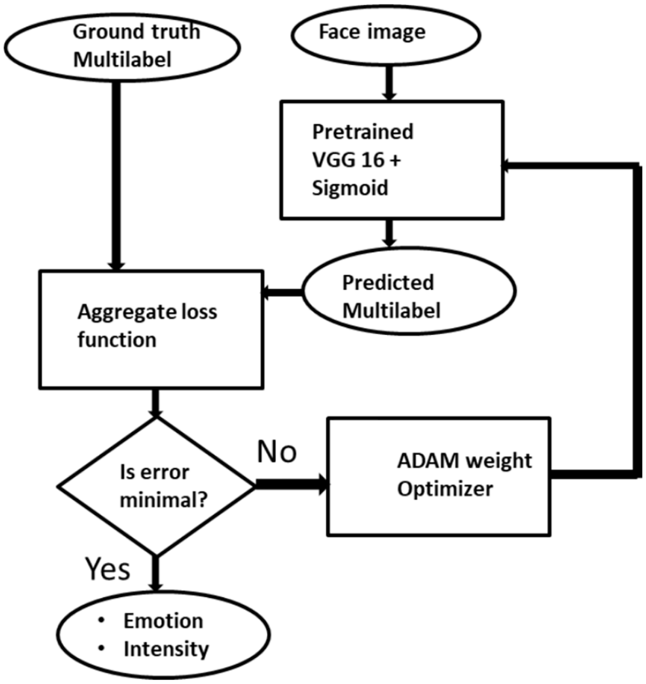


Figure 2 The description of Multilabel CNN model for facial expression recognition and intensity estimation. Full-size DOI: 10.7717/peerj-cs.736/fig-2

aids in adopting deep learning into the multilabel environment. Figure 2 gives the pictorial description of the proposed ML-CNN model.

Convolution neural network multilabel adaptation

The main components of CNN include the convolution layers, the pooling layers, the fully connected layers and the output layer. ML-CNN model is designed similarly with VGG network but with a fewer number of blocks. Figure 3B illustrates the arrangement of all the components of ML-CNN.

Convolution Layer: Convolution layer deals with the extraction of representative features from the expression image; it performs convolution operation on the input image to preserve the spatial relationship between pixels. With convolution operation, local dependencies of the input image are learned. Convolution operation involves convoluting input data with a filter to give a corresponding output which size is determined by some parameters like depth, stride and zero paddings. Convolution layer also employs activation function, which is continuous and differentiable for learning a non-linear transformation of the input data and enhances the network to access a rich hypothesis space from deep representation. This work employs 3 × 3 kernel, ReLu activation function, zero-padding one stride and batch normalisation at each convolution layer. There are five convolution blocks in this model, and the first convolution layer convolutes the input image with the kernel to produce 32 feature maps, a non-linear activation function ReLu is applied to learn the non-linearity features, sparsity control and also to prevent gradient vanishing which is likely to occur during back-propagation. For the stability of each layer,

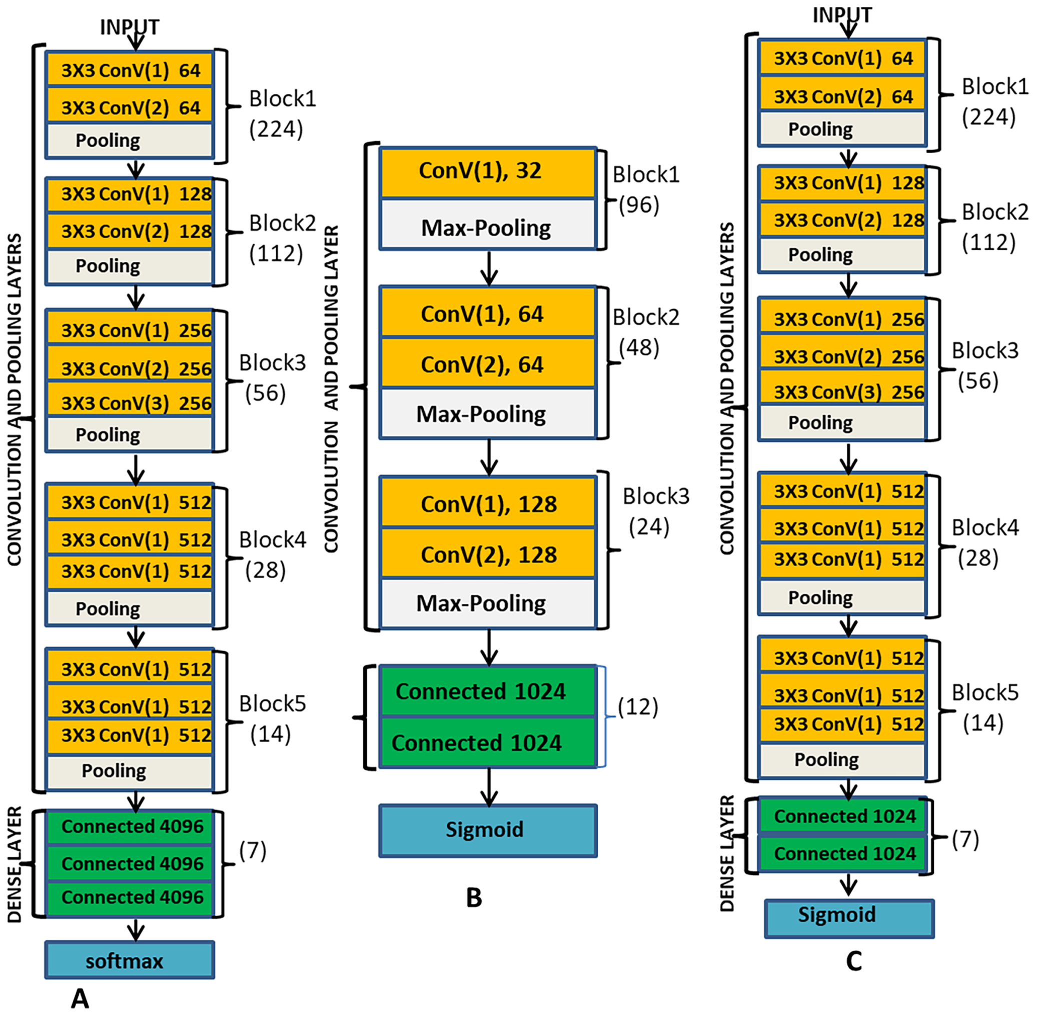


Figure 3 **(A) Description of VGG-16 model; (B) the proposed ML-CNN modeland; (C) the VGGML-CNN model, which the optimised version of ML-CNN.** Full-size DOI: 10.7717/peerj-cs.736/fig-3

we also used batch normalisation and 0.5 dropout. All these operations took place at each of the convolution layers, except that, different filters are generated at other convolution layers. At the second and third convolution layer, 64 feature maps are produced, and at the fifth and sixth layers, 128 feature maps are produced.

Pooling Layer: this is a sub-sampling layer of the network where the down-sampling operation takes place. Its goal is to reduce feature maps' dimension and ensure the preservation of the most useful feature. Pooling operation reduces the computation complexity by reducing the number of training parameters, reducing distortion, and rotation, translation and scaling sensitivity. This system employs max-pooling methods. In the max-pooling feature, maps are convoluted with a 2 × 2 kernel to return the maximum value from each region covered by the kernel. This network contains three pooling layers, and the first pooling layer is positioned after the first convolution layer, the second and the third pooling layers are after the third and the fifth convolution layer respectively as shown in Fig. 3B.

Fully Connected layer: This layer behaves like feed-forward network, the output of the last pooling layer is flattened, that is the 2-dimensional matrix is unrolled into a vector. This is because fully connected layer takes a one-dimensional matrix as input, the flattened function converts the height, the width and the feature maps into a series of feature vectors ($h \times w \times f$). This layer also used ReLu activation function and 0.25 dropout.

Classifier: The last layer of ML-CNN, which is the output layer is a sigmoid classifier, and the sigmoid activation function can generate an independent probability for each of the classes and thus suitable for the multilabel classification task.

Loss Function: Loss function guides the optimisation function in the direction to follow, and it states how close or far is the model prediction to the ground truth. Here, Adaptive Moment (ADAM) optimisation function is considered with learning rate of 0.001. ML-CNN is a multilabel model which implements sigmoid activation function at the output layer, the most appropriate loss function for ML-CNN is Binary Cross-Entropy (BCE) loss. BCE combines the functionality of sigmoid activation function and cross-entropy function in the sense that the loss computes for a class has no effect on the loss computes for other classes, and also form a binary classifier between each of the classes and background class, which is not a member of the classes in consideration. With BCE, loss calculated for each class is independent on the other classes, BCE is formally expressed in (1).

Deep learning networks performance has been enhanced in literature by the modification or introduction of some loss functions like: triplet loss (*Chen et al., 2020*; *Dong & Shen, 2018*; *Vijay Kumar, Carneiro & Reid, 2016*; *Cheng et al., 2016*), center loss (*Wen et al., 2016*), and Island loss (*Cai et al., 2018*). As discussed in *Cai et al. (2018)*, island loss is capable of increasing network discriminating power by increasing interclass variation and reducing intraclass variation. This is the main challenge in FER tasks especially in our model where intraclass variation is large among the representative image samples because each of the classes contains different subjects, and small interclass variation is observed between classes because subjects are the same for all classes. An experiment conducted by *Cai et al. (2018)* showed that island loss function is better in performance than softmax loss function or with center loss function.

$$BCE(s_i) = -\sum_{i=1}^{C=2} t_i log(s_i) \tag{1}$$

where s_i is the model score and t_i is the ground truth for each class $i \in C$.

This work is adapting island loss to enhance the choice of discriminating features in the ML-CNN model. Island loss is an improvement over the center loss with the tendency to minimise or avoid overlapping of different clusters, thus increasing interclass variations. Just like the presentation in *Cai et al. (2018)*, We follow similar steps and positioned island loss function after the fully connected layer. The island loss function is formally expressed in (2).

Algorithm 1 ML-CNN algorithm.

```
Input: Training Data X = {$xi$,($yi \times zi$)}
Output: Network layer parameter W, $L_{IL}$, $L_{BCD}$
Given: minibatch n, learning rate $\alpha$, $\mu$ and hyperparameter $\lambda$ and $\lambda_1$
Initialization: {t, W, $\theta$, $c_j$}
t = 1
while(t != T) {compute the aggregate loss Lagg = $L_{BCE} + \lambda_{LIL}$
update $L_{BCE}$
$\gamma^{t+1} = \gamma^t - \mu(\partial L^t_{BCE})/(\partial \gamma^t)$
update $L_{IL}$
$cj^{t+1} - \alpha\ \Delta cj^t$
update backpropagation error
$\partial L^t/\partial x^t_i = \partial L^t_{BCE}/\partial x^t_i + \lambda(\partial L^t_{IL}/\partial x^t_i)$
Update network layer parameter
$W^{t+1} = W^t - \mu \partial L^t/\partial W^t = W^t - \mu(\partial L^t/\partial x^t_i)(\partial x^t_i/\partial W^t)$
$t = t + 1$}
```

$$\mathcal{L}_{IL} = \mathcal{L}_C + \lambda_1 \sum_{c_j \in N} \sum_{c_k \in N, c_k \neq c_j} \left(\frac{c_k . c_j}{||c_k||_2 ||c_j||_2} + 1 \right) \tag{2}$$

$\mathcal{L}_C$ is the center loss expressed in (3), expression label set is represented with N, both c_k and c_j indicate the two center terms with L_2 norm $||c_k||_2$ and $||c_j||_2$ that penalise the expression of different samples and the similarity of expression samples from the center respectively. λ_1 is to balance c_k and c_j

$$\mathcal{L}_c = \frac{1}{2} \sum_{i=1}^{m} ||x_i - c_{yi}||^2 \tag{3}$$

ML-CNN model implements BCE loss function at the final layer, then the entire loss function of ML-CNN is provided in (4).

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{IL} \tag{4}$$

where $\mathcal{L}_{BCE}$ is Binary Cross Entropy loss, and λ is a hyper-parameter for balancing the two losses. The implementation procedure of ML-CNN is detail in Algorithm 1.

Transfer learning

Transfer Learning could be thought of as a way of preventing re-inventing the wheels in computer vision, in the sense that knowledge of a particular deep model could be transferred or reuse more especially in a similar environment or for a related task. Transfer learning mechanism improvises for data challenges in computer vision, and it is considered as one of the deep learning optimisation techniques for addressing overfitting effect in the field. Adapting the knowledge or weight of pre-trained standard deep network into a related task or challenge is the main concept of transfer learning. Few of the standard deep pre-trained networks include: VGG Network (VGGNET) (*Simonyan & Zisserman, 2015*), Residual Network (ResNet) (*He et al., 2016*), Inception_w (*Szegedy et al., 2017*; *Szegedy et al., 2016*), Google Network (GoogLENet) (*Szegedy et al., 2015*) and the likes.

ML-CNN design is a similitude of VGG; we consider VGG-16 as a pre-trained network for our model. The pre-trained network is adapted as a feature extractor for our ML-CNN. The fully connected layers and the ML-CNN classifier control the learning and the interpretation of the extracted features on the datasets and preserve both the multilabel learning and independent classification. Figure 3 is the pictorial description of the VGGML-CNN model. Figure 3A is the description of VGG-16 model, Fig. 3B is the proposed ML-CNN model, and Fig. 3C is the VGGML-CNN model, the optimised version of ML-CNN.

EXPERIMENT

Preprocessing

Deep learning is known for its autonomous feature extraction capability. Despite, observations show that there is an improvement in networks performance when data is preprocessed. Data preprocessing advantages to deep learning include minimization of computational costs (computational time and computational resources) and availability of proper representative features that is noise-free. In this work, we find it appropriate to employ some preprocessing techniques to aid our model sensitivity in the automatic feature extraction phase. This section carried out two essential data preprocessing techniques: face localization (face detection) and face augmentation.

Face localization

Face detection is about locating the region of a face from an image, sequence of images or video. Face detection algorithms are often involved in virtually most face related research in computer vision such as face recognition, Age estimation from face, image-based gender recognition and facial expression recognition. All these tasks consider face detection as one of the main steps in their preprocessing stage. In this study, we consider the face detection algorithm proposed by *Viola & Jones (2001)*. The only modification to this algorithm is the implementation of an integral graph for eigenvalues computation as in *Zhang, Jolfaei & Alazab (2019)* which aid the computation speed, we use the method to compute Haar-like feature *via* integral graph as shown in Eq. (5). In the process, relevant features of Haar-like are carefully selected and later integrated into a robust classifier with the aid of the AdaBoost algorithm.

$$G(x,y)I(x,y) = \sum_{x' \leq x, y'} i(x', y') \tag{5}$$

where I(.) is the integral image and i(.) is the real image.

Augmentation

Augmentation is one of the policies employed in computer vision to alleviate data limitation challenge. Data augmentation is mostly used in deep learning where there is a need for extensive data size for deep learning model to learn the representative feature appropriately from the data sample when training the network. Data augmentation could be implemented on the fly or off the fly. This work implements off the fly techniques using the augmentor module in python3 for data balancing among the classes.

Experimental databases

(1) Binghamton University 3D Facial Expression (BU-3DFE): BU-3DFE (*Yin et al., 2006*) is a controlled static dataset that captured real-world challenges. It consists of 2,00 data collected form 100 subjects. Each of the subjects produced four images for each of the six basic emotion classes (Anger, Disgust, Fear, Happy, Sadness, Surprise) with their respective intensity annotated with ordinal metrics. BU-3DFE is the only FER dataset that considers ordinal intensity annotation in the database to the best of our knowledge.

(2) Cohn Kanade Extention (CK+): CK+ (*Lucey et al., 2010*) is a sequence dataset and well-annotated into seven basic expression classes (Anger, Disgust, Contempt, Fear, Happy, Sadness and Surprise). It is made up of 327 sequence data collected from 118 subjects. A subject produced an emotion sequence for each of the seven basic emotions starting from the neutral face (offset) to the onset and apex. CK+ is a popular dataset for facial intensity estimation, and for this study, the data is organised following the flow of changes in the sequence to have an ordinal label in substitute for the onset, offset and apex. The sequence of expression for each subject is categorised into four ordinal intensities {Low, Normal, High, and V High} according to the observed changes. This implies that each emotion will have four sub-classes tagged with the emotion and each of the ordinal intensities. For instance, a subject with an anger expression sequence would be grouped into AngerLow, AngerNormal, AngerHigh and AngerVery High. The arrangement makes CK conform to the ordinal intensity arrangement in BU-3DFE datasets.

Experiment procedures

This section evaluates the proposed ML-CNN model and the comparative study of its performance with the existing multilabel models. BU-3DFE and CK+ data are the set of databases employed for the experiments. After pre-processing, each of the raw data was scaled to a uniform size of 96 × 96. The pixel values were divided by 255 to ensure data scale normalisation. The datasets are partitioned into the training set (70%), the validation set (20%), and the remaining 10% is the testing set. The experiment was conducted using OpenCV, Scikit-learn, Keras with TensorFlow 2.0 backend. All the required software were installed on High-Performance Computing (HPC) hardware resources at the Center for High-Performance Computing (CHPC).

Evaluation of ML-CNN model begins with training procedure. The model was first trained on the training data division and evaluated on the validating data severally with some modifications to the model parameters to minimise the model's over-fitting. Adam optimiser with initial learning rate of 0.001 is used. Initially, we consider the model's performance evaluation on the BU-3DFE data with a data size of 2,400. We also extend the experiment to observed the system performance when the training data is augmented. The augmentation is implemented offline to ensure data balance among the classes. Here, we evaluate the system performance on both BU-3DFE and CK+ datasets. We also observed the transfer learning optimisation technique on the model by fine-tuning the model with a pre-trained VGG-16 CNN network model. We employ accuracy and the loss (binary cross entropy and island loss) metrics for the model performance evaluation on the testing data in each of the described experiment.

The other phase of our experiment is a comparative study of the ML-CNN and four different other multi-label algorithms: RAKELD (Distinct Random k-Label sets) (*Tsoumakas, Katakis & Vlahavas, 2011*), classifier chain (CC) (*Read et al., 2009*), MLkNN (Multilabel k Nearest Neighbour) (*Zhang & Zhou, 2007*) and MLARAM (*Benites & Sapozhnikova, 2015*). To avoid bias, the algorithms were implemented in the same environment and executed on similar datasets with fair consideration by using multilabel performance evaluation metrics. Gaussian Naive Bayes is the based classifier in RAKELD, the base classifier for CC is the random forest, while kNN is used as the base classifier for MLkNN nearest neighbour k is set to 10, and smoothing parameter is 1. The multilabel metrics used for our models' comparative studies with other models include average precision, hamming loss, coverage error, and ranking loss. The following section contains a brief discussion of each of the listed multilabel metrics.

Evaluation metrics

Hamming Loss: is computed using the XOR operator as the loss between the predicted and actual labels. The Hamming loss is defined in Eq. (6).

$$H = \frac{1}{|N|.|L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|J|} XOR(y_{i,j}, \hat{y}_{i,j}) \tag{6}$$

Ranking Loss: computes the average of the incorrectly ordered labels. The smaller the Ranking loss, the better the performance of the model. The Ranking loss is defined in Eq. (7).

$$Rank_{loss}(y, \hat{f}) = \frac{1}{N} \sum_{1=0}^{N-1} ||y_i||_0 \frac{1}{k - ||y_i||_0} |Z| \tag{7}$$

where k is the number of labels and Z is (m,n): $\hat{f}_{i,m} \geq \hat{f}_{i,n}$, $y_{i,m} = 1$, $y_{i,n} = 0$

Average Precision: is the number of higher-ranked labels that are true for each ground-truth label. The higher the Average precision value, the better the performance of the model. The ranking average precision is defined in Eq. (8).

Table 2 The tabular presentation of ML-CNN and VGGML-CNN performance evaluation Using accuracy and aggregate loss on BU-3DFE and CK+ datasets, and their comparison with some existing methods. In the table, metric with ↑ indicates that the higher the metric value the better the model performance, and metric with ↓ indicates that the lower or smaller the value of the metric the better the model performance.

ML-Models	Database	Accuracy ↑	Aggregate loss ↓
ML-CNN	BU-3DFE	88.56	0.3534
	AUG_BU-3DFE	92.84	0.1841
	CK+	93.24	0.2513
VGGML-CNN	Bu-3DFE	94.18	0.1723
	AUG_BU-3DFE	98.01	0.1411
	CK+	97.16	0.1842
Kamarol et al. (2017)	CK+	82.4	NA
Walecki et al. (2015)	CK+	94.5	NA
Quan, Qian & Ren (2014)	CK+	88.3	NA

$$LRAP = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{||y||_0}\sum_{j:y_{i,j}=1}\frac{|L_{i,j}|}{Ri,j} \tag{8}$$

where $L_{i,j}$ = K: $y_{i,k}$=1,$\hat{f}_{i,k} \geq \hat{f}_{i,j}$ and |.| is the cardinality of the set.

Coverage Error: computes the number of labels required to included all the correct labels in the final prediction. The smaller the value, the better the model performance of the model. The Coverage error is defined in Eq. (9).

$$Coverage(y,\hat{f}) = \frac{1}{N}\sum_{i=0}^{N-1} maxrank_{i,j} \tag{9}$$

where $rank_{i,j}$ is—{k:$\hat{f}_{i,k} \geq f_{i,k}$—}

In addition to the comparative studies, we visually observed the model prediction output and compared the degree of intensity predicted for each expression with the truth label.

EXPERIMENTAL RESULTS AND DISCUSSION

The experiments' results are summarised in the figures and the tables below. The experiments first observe ML-CNN model's performance and the optimisation technique adopted on both the BU-3DFE and CK+ datasets. We use accuracy and the loss function as the model evaluation metrics. Observations showed that ML-CNN based model provides a training accuracy of 95% and validation accuracy of 88.56%, training loss and validation loss of 0.142 and 0.3534, respectively. Augmentation of training data improves ML-CNN performance with about 2% increase in training accuracy and almost 4% increase in validation accuracy. The result obtained by fine-tuning ML-CNN with VGG network improves the model performance with validation accuracy close to 8%. The summary of these results is presented in Table 2. Table 2 shows that our model

Table 3 **The result of the comparative studies of multilabel models' performances on BU-3DFE dataset is presented as follows.** Metric with ↑ indicates the higher the metric value, the better the model performance, and metric with ↓ indicates the lower or smaller the value of the metric the better the model's performance.

ML-Models	Hamming loss ↓	Ranking loss ↓	Average precision ↑	Coverage ↓
RAKELD	0.4126	0.6859	0.2274	4.8137
CC	0.1807	0.8393	0.3107	4.8094
MLkNN	0.1931	0.8917	0.2634	4.9486
MLARAM	0.3045	0.6552	0.3180	3.1970
ML-CNN	0.1273	0.2867	0.5803	2.5620
VGGML-CNN	0.0890	0.1647	0.7093	1.9091

Table 4 **The comparative studies of multilabel models' performances on augmented BU-3DFE dataset are presented as follows.** Metric with ↑ indicates the higher the metric value, the better the model performance, and metric with ↓ indicates the lower or smaller the value of the metric the better the model's performance.

ML-Model	Hamming loss ↓	Ranking loss ↓	Average precision ↑	Coverage ↓
RAKELD	0.3858	0.7223	0.2241	4.0453
CC	0.1825	0.8948	0.2812	4.7270
MLkNN	0.1929	0.9025	0.2573	4.9623
MLARAM	0.3169	0.6963	0.3280	2.9315
ML-CNN	0.1124	0.2278	0.7216	2.2397
VGGML-CNN	0.0628	0.1561	0.8637	1.3140

outperforms some existing methods on facial expression recognition and intensity estimation. Although the methods in consideration either recognised expression before the intensity estimation or model the intensity estimation before expression recognition, which is quite different from our model that recognises expression and intensity concurrently.

The outcomes of our comparative studies of ML-CNN models with some other multilabel algorithms are presented in Tables 3–5. It is evident from the tables that our proposed multilabel model shows a better performance than the multilabel algorithms considered. Table 3 indicates that ML-CNN and VGGML-CNN give outstanding performances over RAKELD, CC, MLkNN and MLARAM when predicting emotion and the degree of intensity BU-3DFE. Observation from Table 4 clearly showed that RAKELD, CC, MLkNN and MLARAM degrade in performance on Augmented BU-3DFE data, unlike ML-CNN VGGML-CNN that showed significant improvement in their performance under similar conditions. Table 5 also shows that both VGGML-CNN and ML-CNN outperformed other multilabel algorithms. Furthermore, Tables 6 and 7 contain the detail predictions of each expression and intensity on the test samples of the datasets. Figures 4 and 5 presents the multilabel confusion matrix for VGGML-CNN performance

Table 5 **The result of the comparative studies of multilabel models' performances on CK+ dataset is presented as follows.** Metric with ↑ indicates the higher the metric value, the better the model performance, and metric with ↓ indicates the lower or smaller the value of the metric the better the model's performance.

ML-Model	Hamming loss ↓	Ranking loss ↓	Average precision ↑	Coverage ↓
RAKELD	0.3904	0.6637	0.2370	4.4435
CC	0.1489	0.6842	0.4234	4.7339
MLkNN	0.1839	0.8345	0.2965	4.7930
MLARAM	0.1951	0.4636	0.4144	3.0748
ML-CNN	0.1487	0.4161	0.5926	2.8120
VGGML-CNN	0.1393	0.3897	0.6002	1.4359

Table 6 **Emotion and intensity degree predictions on BU-3DFE test samples.**

EMotion and ordinal intensity	Accuracy %
Anger	97.0
Disgust	98.3
Fear	97.0
Happy	100
Sadness	98.7
Surprise	98.7
Low	98.7
Normal	97.5
High	97.5
Very High	97.0

Table 7 **Emotion and intensity degree prediction on CK+ test samples.**

Emotion and ordinal intensity	Accuracy %
Anger	98.1
Disgust	98.1
Fear	100
Happy	98.1
Sadness	100
Surprise	100
Low	96.2
Normal	83.3
High	87.0
Very High	96.3

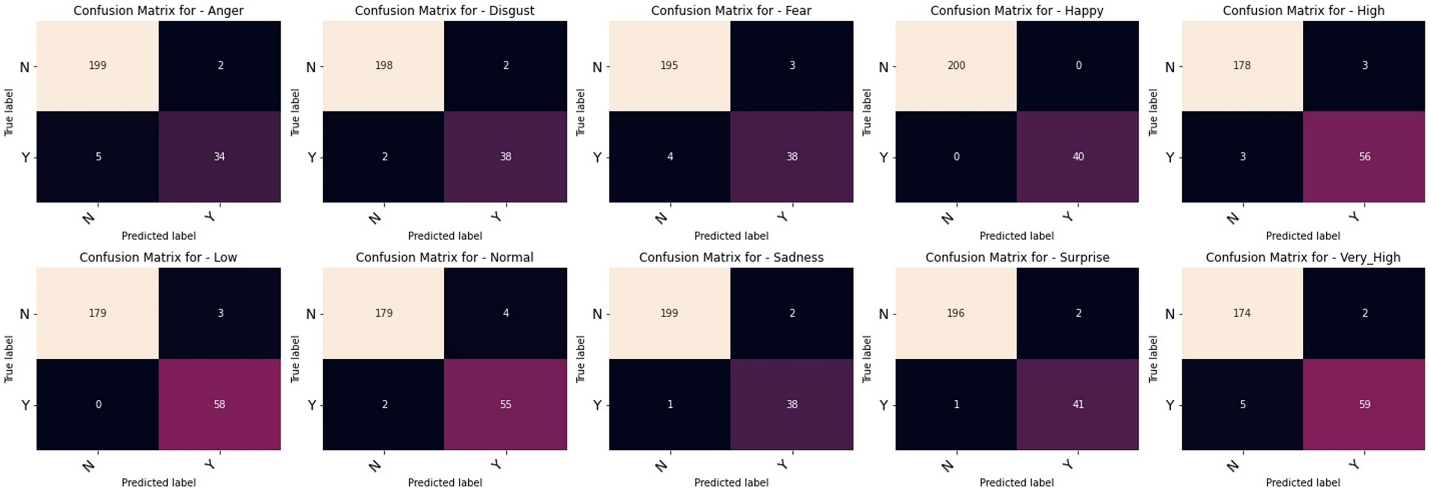


Figure 4 **Multilabel confusion matrix of the VGGML-CNN on Bu-3DFE.** Full-size DOI: 10.7717/peerj-cs.736/fig-4

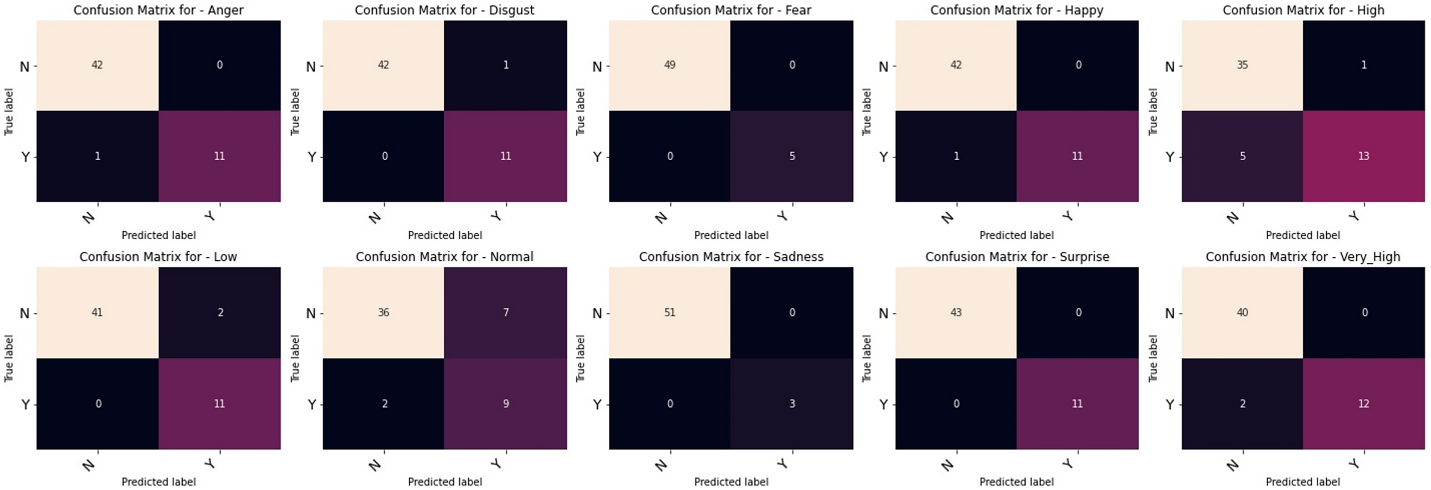


Figure 5 **Multilabel confusion matrix of the VGGML-CNN on CK+.** Full-size DOI: 10.7717/peerj-cs.736/fig-5

on both the BU-3DFE and CK+ respectively. VGGML-CNN performance is compared with some of the recent models of FER using CK+ and BU-3DFE datasets. VGGML-CNN shows outstanding performance on BU-3DFE, and a good result on CK+, detail of the comparative study is presented in Tables 8 and 9.

Equity and model bias

Although BU-3DFE is static data, also regarded as data-in-the-wild, the data comprises of subjects of different ages, ethnicity, races, and genders. Other factors that possibly challenge FER recognition are considered in the collection of the data. The result of the model on BU-3DFE shows that human variation factors have limited effect on the model.

Table 8 **Comparison result of VGGML-CNN with some recent models on CK+.**

Model	Accuracy %	No of classes	Target
Cai et al. (2018) (IL-CNN)	94.35	7	Expression only
Li & Deng (2019) (DLP-CNN)	95.78	7	Expression only
Alenazy & Alqahtani (2020) (DBN-GSA)	98	7	Expression only
Xu et al. (2020) (CCNN)	91.50	6	Expression and intensity
Chen et al. (2020) LDL-ALSG	93.08	7	Expression distribution
ML-CNN	93.24	6	Expression and intensity
VGGML-CNN	97.16	6	Expression and intensity

Table 9 **Comparison result of VGGML-CNN with some recent models on BU-3DFE.**

Model	Accuracy %	No of classes	Target
Fernandez et al. (2019) (FERAtt)	85.15	7	Expression only
Shao & Qian (2020) (MVFE-LightNet + Residual Convolution)	88.70	6	Expression only
Bao, Zhao & Chen (2020) (CNM)	80.63	6	Expression only
VGGML-CNN	98.01	6	Expression and intensity

CONCLUSIONS

This work proposed a new approach to FER and intensity estimation. The multilabel convolution neural network (ML-CNN) method employed problem transformation technique and used CNN as the binary classifier to predict emotion and its corresponding intensity estimation using ordinal metrics. For system robustness and accuracy reliability, we used transfer learning optimisation as a trade-off for the small data population and overfitting prevention. We modified the loss function by introducing island loss function to enhance the model sensitivity to intraclass variation minimisation and interclass variation maximisation. Our proposed model accurately predicts the emotional state with the corresponding degree of intensity concurrently. From the comparative study of ML-CNN with other multilabel algorithms, ML-CNN shows significant performance advantage, more especially with both augmented data and the model optimisation. Despite the excellent performance, the ML-CNN model still finds it difficult to generalise to unseen data outside the scope of the databases used. We suspect infiltration of person specificity, that is, personal identity into the training time model, as the possible reason. The drawback should be considered in the future work for a robust ML-CNN model that will generalise well with unseen data. In addition, future work should also consider using some spontaneous data-in-the-wild that have no hierarchical intensity organization, such as FER2013 and FER+, and dynamic FER data that would support the real-life application of the model.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Olufisayo Ekundayo conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Serestina Viriri conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data we used are available at:

- BU-3DFE: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html

- CK+: CK+ is available upon approved request: https://www.jeffcohn.net/resources/

Our Python code is available as a Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.736#supplemental-information.

REFERENCES

Aamir M, Ali T, Shaf A, Irfan M, Saleem MQ. 2020. ML-DCNNet: multi-level deep convolutional neural network for facial expression recognition and intensity estimation. *Arabian Journal for Science and Engineering* **45(12)**:10605–10620 DOI 10.1007/s13369-020-04811-0.

Alenazy WM, Alqahtani AS. 2020. Gravitational search algorithm based optimized deep learning model with diverse set of features for facial expression recognition. *Journal of Ambient Intelligence and Humanized Computing* **12(1631)**:1646 DOI 10.1007/s12652-020-02235-0.

Bao W, Zhao Y, Chen D. 2020. A facial expression recognition method using capsule network model. *Scientific Programming* **2020**:805–814 DOI 10.1155/2020/8845176.

Batista JC, Albiero V, Bellon ORP, Silva L. 2017. AUMPNet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. In: *Proceedings-12th IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2017-1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017*. Heteroge. Piscataway: IEEE, 866–871.

Behere RV. 2015. Facial emotion recognition deficits: the new face of schizophrenia. *Indian Journal of Psychiatry* **57(3)**:229–235 DOI 10.4103/0019-5545.166641.

Benites F, Sapozhnikova E. 2015. Haram: a hierarchical aram neural network for large-scale text classification. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. Piscataway: IEEE, 847–854.

Cai J, Meng Z, Khan AS, Li Z, Oreilly J, Tong Y. 2018. Island loss for learning discriminative features in facial expression recognition. In: *Proceedings-13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. Xi'an: IEEE, 302–309.

Chang KY, Chen CS, Hung YP. 2013. Intensity rank estimation of facial expressions based on a single image. In: *2013 IEEE International Conference on Systems, Man, and Cybernetics*. Manchester: IEEE, 3157–3162.

Chen Z, Ansari R, Wilkie D. 2012. Automated detection of pain from facial expressions: a rule-based approach using AAM. *Progress in Biomedical Optics and Imaging-Proceedings of SPIE* **8314**:125 DOI 10.1117/12.912537.

Chen S, Wang J, Chen Y, Shi Z, Geng X, Rui Y. 2020. Label distribution learning on auxiliary label space graphs for facial expression recognition. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 13981–13990.

Cheng D, Gong Y, Zhou S, Wang J, Zheng N. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 1335–1344.

Dong X, Shen J. 2018. Triplet loss in Siamese network for object tracking. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 472–488.

Du S, Martinez A. 2015. Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in Clinical Neuroscience* **17**:443–455 DOI 10.31887/dcns.2015.17.4/sdu.

Ekman P, Friesen WV. 1971. Constant across cultures in the face and emotion. *Journal of Personality and Social Psychology* **17(2)**:124–129 DOI 10.1037/h0030377.

Ekundayo Olufisayo, Viriri Serestina. 2019. Facial expression recognition: a review of methods, performances and limitations. In: *2019 Conference on Information Communications Technology and Society, ICTAS 2019*. Piscataway: IEEE, 1–6.

Ekundayo O, Viriris S. 2020. *Facial expression recognition and ordinal intensity estimation: a multilabel learning approach*. Vol. 12510. Cham: Springer.

Fan Y, Lu X, Li D, Liu Y. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: *ICMI, 2016-Proceedings of the 18th ACM International Conference on Multimodal Interaction*. New York: ACM, 445–450.

Fernandez PM, Penã FAG, Ren TI, Cunha A. 2019. FERAtt: facial expression recognition with attention net. *Available at https://arxiv.org/abs/1902.03284*.

Gudi A, Tasli HE, den Uyl TM, Maroulis A. 2015. Deep learning based FACS action unit occurrence and intensity estimation. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*. Piscataway: IEEE.

Guo W, Yang H, Liu Z, Xu Y, Hu B. 2021. Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Frontiers in Neuroscience* **15**:342 DOI 10.3389/fnins.2021.609760.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 770–778.

Kahou SE, Michalski V, Konda K, Memisevic R, Pal C. 2015. Recurrent neural networks for emotion recognition in video. In: *ICMI, 2015-Proceedings of the 2015 ACM International Conference on Multimodal Interaction*. New York: ACM, 467–474.

Kamarol SKA, Jaward MH, Kälviäinen H, Parkkinen J, Parthiban R. 2017. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. *Pattern Recognition Letters* **92(10)**:25–32 DOI 10.1016/j.patrec.2017.04.003.

Kim M, Pavlovic V. 2010. Hidden conditional ordinal random fields for sequence classification. *ECML PKDD* **6322**:51–65 DOI 10.1007/978-3-642-15883-4.

Lee KK, Xu Y. 2003. Real-time estimation of facial expression intensity. In: *Proceedings-IEEE International Conference on Robotics and Automation*. Taipei: IEEE, 2567–2572.

Li S, Deng W. 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* **28(1)**:356–370 DOI 10.1109/TIP.2018.2868382.

Li S, Deng W. 2020. Deep facial expression recognition: a survey. *Available at https://arxiv.org/abs/1804.08348*.

Liu Y, Cao Y, Li Y, Liu M, Song R, Wang Y, Xu Z, Ma X. 2016. Facial expression recognition with PCA and LBP features extracting from active facial patches. In: *2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016*. Angkor Wat: IEEE, 368–373.

Liu X, Vijaya Kumar BVK, You J, Jia P. 2017. Adaptive deep metric learning for identity-aware facial expression recognition Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Piscataway: IEEE, 20–29.

Liu W, Wen Y, Yu Z, Yang M. 2016. Large-margin Softmax loss for convolutional neural networks. In: *Proceedings of The 33rd International Conference on Machine Learning, ICML 2016. PMLR*. 507–516.

Luaces O, Díez J, Barranquero J, del Coz J, Bahamonde A. 2012. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence* **1(4)**:303–313 DOI 10.1007/s13748-012-0030-x.

Lucey P, Cohn JF, 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. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, CVPRW 2010*. San Francisco: IEEE, 94–101.

Mollahosseini A, Hasani B, Mahoor MH. 2019. Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* **10(1)**:18–31 DOI 10.1109/TAFFC.2017.2740923.

Nomiya H, Sakaue S, Hochin T. 2016. Recognition and intensity estimation of facial expression using ensemble classifiers. *International Journal of Networked and Distributed Computing* **4(4)**:203–211 DOI 10.2991/ijndc.2016.4.4.1.

Ozcan T, Basturk A. 2020. Static facial expression recognition using convolutional neural networks based on transfer learning and hyperparameter optimization. *Multimedia Tools and Applications* **79(35–36)**:26587–26604 DOI 10.1007/s11042-020-09268-9.

Plutchik R. 2001. Integration differentiation and derivatives of emotion. *Evolution and Cognition*. **7(2)**:114–125.

Quan C, Qian Y, Ren F. 2014. Dynamic facial expression recognition based on K-order emotional intensity model. In: *2014 IEEE International Conference on Robotics and Biomimetics, IEEE ROBIO 2014*. Piscataway: IEEE, 1164–1168.

Read J, Pfahringer B, Holmes G, Frank E. 2009. Classifier chains for multi-label classification. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J, eds. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science*. Vol. 5782. Berlin, Heidelberg: Springer, 254–269 DOI 10.1007/978-3-642-04174-7_17.

Roy SD, Bhowmik MK, Saha P, Ghosh AK. 2016. An approach for automatic pain detection through facial expression. *Procedia Computer Science* **84(7889)**:99–106 DOI 10.1016/j.procs.2016.04.072.

Rudovic O, Pavlovic V, Pantic M. 2012. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Providence: IEEE.

Russell JA, Lanius UF. 1984. Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology* **4(2)**:119–135 DOI 10.1016/S0272-4944(84)80029-8.

Seo E, Park HY, Park K, Koo SJ, Lee SY, Min JE, Lee E, An SK. 2020. Impaired facial emotion recognition in individuals at ultra-high risk for psychosis and associations with schizotypy and paranoia level. *Frontiers in Psychiatry* **11**:577 DOI 10.3389/fpsyt.2020.00577.

Shao J, Qian Y. 2020. Multi-view facial expression recognition with multi-view facial expression light weight network. *Pattern Recognition and Image Analysis* **30(4)**:805–814.

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations, ICLR, 2015 - Conference Track Proceedings*. 1–14.

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: *31st AAAI Conference on Artificial Intelligence, AAAI 2017*. 4278–4284.

Szegedy C, Liu W, Jia Y, Reed S, Sermanet P, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 1–12.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2818–2826.

Tsoumakas G, Katakis I, Vlahavas I. 2011. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering* **23(7)**:1079–1089 DOI 10.1109/TKDE.2010.164.

Turan C, Lam K-M. 2018. Histogram-based local descriptors for facial expression recognition (FER): a comprehensive study. *Journal of Visual Communication and Image Representation* **55(January)**:331–341 DOI 10.1016/j.jvcir.2018.05.024.

Valstar MF, Pantic M. 2012. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transaction on Systems, Man, And Cybernetics-Part B: Cybernetics* **42(1)**:28–43 DOI 10.1109/TSMCB.2011.2163710.

Verma R, Davatzikos C, Loughead J, Indersmitten T, Hu R, Kohler C, Gur RE, Gur RC. 2005. Quantification of facial expressions using high-dimensional shape transformations. *Journal of Neuroscience Methods* **141(1)**:61–73 DOI 10.1016/j.jneumeth.2004.05.016.

Vijay Kumar BG, Carneiro G, Reid I. 2016. Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. 5385–5394.

Viola P, Jones M. 2001. Rapid object detection using a boosted cascade of simple features. In: *Conference on Computer Vision and Pattern Recognition 2001*. 1–9.

Walecki R, Rudovic O, Pavlovic V, Pantic M. 2015. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*. Ljubljana: IEEE.

Wang F, Cheng J, Liu W, Liu H. 2018. Additive margin Softmax for face verification. *IEEE Signal Processing Letters* **25(7)**:926–930 DOI 10.1109/LSP.2018.2822810.

Wen Y, Zhang K, Li Z, Qiao Y. 2016. A discriminative feature learning approach for deep face recognition. *Lecture Notes in Computer Science* **9911**:499–515 DOI 10.1007/978-3-319-46478-7.

Wu J, Wang Y, Wang Y. 2019. Matrix transformation-based optimized CNN. In: *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. Dalian: IEEE Explore, 845–849.

Xu R, Chen J, Han J, Tan L, Xu L. 2020. Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. *Computing* **102(3)**:765–780 DOI 10.1007/s00607-019-00722-7.

Yang Y, Sun Y. 2017. Facial expression recognition based on arousal-valence emotion model and deep learning method. In: *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*. 59–62.

Yannakakis GN, Cowie R, Busso C. 2017. The ordinal nature of emotions. In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 248–255.

Yin L, Wei X, Sun Y, Wang J, Rosato MJ. 2006. A 3D facial expression database for facial behavior research. In: *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*. 211–216.

Zatarain Cabada R, Rodriguez Rangel H, Barron Estrada ML, Cardenas Lopez HM. 2020. Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems. *Soft Computing* **24(10)**:7593–7602 DOI 10.1007/s00500-019-04387-4.

Zhang H, Jolfaei A, Alazab M. 2019. A face emotion recognition method using convolutional neural network and image edge computing. *IEEE Access* **7**:159081–159089 DOI 10.1109/ACCESS.2019.2949741.

Zhang M-L, Zhou Z-H. 2007. Ml-knn: a lazy learning approach to multi-label learning. *Pattern Recognition* **40(7)**:2038–2048 DOI 10.1016/j.patcog.2006.12.019.

Zhao R, Gan Q, Wang S, Ji Q. 2016. Facial expression intensity estimation using ordinal information. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 3466–3474.

Zhou Y, Pi J, Shi BE. 2017. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In: *Proceedings-12th IEEE International Conference on Automatic Face and Gesture Recognition, FG, 2017-1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*. Piscataway: IEEE, 872–877.

3.4.2 Conclusion

This paper presented the efficiency of the proposed framework for simultaneous expression recognition and intensity estimation with ordinal metrics. The high performance reported on CK+ and BU-3DFE could be attributed to the preprocessing techniques, the pretrained network and the island loss functions employed.

3.5 Facial Expression Recognition using Manifold Learning and Graph Convolutional Network

3.5.1 Introduction

This section presents a framework that resolves facial expression label ambiguity and data annotation inconsistency by learning the correlation among labels with manifold learning that models labels correlations with similarity distance and employs graph convolutional network in a semisupervised manner to recover the distribution of emotions from logical labels. Facial expression is described as a graph related problem where the facial expression data point forms the graph's nodes, and similarity distances produced by manifold learning form the graph's edge information.

ORIGINAL ARTICLE

Journal Section

Facial expression recognition with manifold learning and graph convolutional network

Olufisayo Ekundayo[1*] | Serestina Viriri[1*]

[1]School of Mathematics,Statistics and Computer Science, University of KwaZulu-Natal, Durban, KwaZulu-Natal, 4000, South Africa

Correspondence
School of Mathematics,Statistics and Computer Science, University of KwaZulu-Natal, Durban, KwaZulu-Natal, 4000, South Africa
Email: viriris@ukzn.ac.za

Funding information

Facial Expression Recognition (FER) has the ability to detect human affect state. Most of the methods employed for FER task do not really consider the correlation among FER data labels to resolve data annotation and ambiguity problems. Label Distribution Learning (LDL) application to FER considerably address this, but only in the presence of data with distribution labels. Therefore, methods that would recover label distribution from logical labels are required. This work is presenting a graph-based label enhancement approach with manifold learning and Graph Convolutional Network (GCN) for facial expression recognition. The manifold learning approach transforms FER data as a graphical problem, where the data points are considered as nearest neighbours represent graph nodes, with the motive of representing the distances along the edges of the neighbouring graph with the approximate distances along the manifold. This process uses the nearest neighbour graph to learn the geometric structure in FER data, which also learn the possible correlation among the data labels. The graphical convolutional network is employed to incorporate the information provided in the manifold learning and the logical description of the data to classify the nodes of the graph using the information of the nearest neighbours. The experiment conducted

[*] Equally contributing authors.

on the Binghampton University-3D Facial Expression (BU-3DFE) and the Cohn Kanade extension (CK+) data shows that the model gives promising results.

KEYWORDS
Facial Expression Recognition, Manifold Learning, Graph Convolutional Network, Label Distribution Learning, Label Enhancement.

1 | INTRODUCTION

As one of the prominent research fields in cognitive computing and Computer Vision (CV), FER has produced positive and promising research results that encourage its public acceptance and applications. FER application list is not limited to medicine, security, marketing and education. Recently, Liu et al. (2021) presented a work that captured student emotional state in real-time in a virtual class. They incorporate emotion recognition into virtual learning. The successes recorded in FER so far could be attributed to the extensive research on developing robust FER models and the quality and quantity of available FER datasets. FER could be divided into three main aspects: the data, the model and the annotation, which is the description of the data classes. In the literature, little consideration is given to the data annotation aspect. The prevailing data annotation in FER is the grouping of FER data into logical labels of six or seven classes, as proposed by Ekman and Friesel (Ekman and Friesen, 1971). Nevertheless, the logical label of FER datasets requires the hands of experts, which is expensive and time-consuming. Also, human errors like annotation inconsistency and bias are probably inevitable due to the ambiguous nature of FER datasets. Another shortcoming with the logical labels is its inability to consider the correlation among labels.

Label distribution learning and label enhancement techniques have been employed in the literature to mitigate the ambiguity nature and the label correlation challenges in FER Jia et al. (2019) Chen et al. (2020). Among the label distribution enhancement algorithms, graph-based models have reported promising results. The graph-based model for label enhancement relies on the assumption that data features and data labels have smooth variations along the graph's edges. While label propagation algorithms aggregate node labels over the edges in the graph, the manifold learning algorithms present distance along the edges with the approximate distance along the manifold. Chen et al. (2020) reported that the graph algorithms that have a strong local linear assumption and strong smoothness assumptions are not suitable for image input and deep neural features. But Isomap manifold is global manifold learning to preserve geometric features of data in a lower-dimensional space Xie et al. (2017).

This work presents a graph-based label enhancement approach with manifold learning and Graph Convolutional Network (GCN) for facial expression recognition. The manifold learning approach transforms FER data as a graphical problem, where the data samples form the graph nodes, and the edges of the neighbouring graph are represented with the approximate distances along the isomeric manifold. The Isomap employs the image-based Euclidean distance and geodesic algorithm to learn the geometric structure and the correlation among data annotations. GCN propagates data features along the edges of the neighbouring nodes, incorporates the information provided in the manifold learning and the logical description of the data, and classifies the graph's nodes into six basic emotions using a semi-supervised approach.

Our contributions include the following:

- Incorporate inductive and transductive learning to model the correlation among facial expression labels and recover emotion distributions from logical labels.
- The Isomap manifold learning is employed to learn the correlation among FER data annotations by computing the similarity distance of neighbouring nodes. The correlation is modelled with the objectives that nodes closed in the manifold are likely to have similar annotation.
- Graph Convolution Network is used in a semi-supervised manner to propagate data features along the edges of the neighbouring nodes and thus learn emotion distribution from the neighbouring nodes.

The proposed method is similar to the method in Chen et al. (2020) and He and Jin (2019) by using a manifold algorithm for data similarity computation. However, our approach is different from the work of Chen et al. (2020) by using Isomap manifold, which preserves the global representative features in the data and also minimises parameter tuning challenge. The main parameter is K, which is the number of neighbouring nodes that determines the topological stability. Also, the proposed model is different from the work of He and Jin (2019) by computing image-based Euclidean distance that makes the model robust against short-circuit error caused by general Euclidean distance used in He and Jin (2019).

This work is organised as follows: Section 2 contains the review of some recent works in FER that considered label annotation challenges in FER model; Section 3 presents a discussion of the proposed model starting from Isomap manifold to GCN model. A detail account of the experiments is presented in Section 4, which provides information about the datasets, data preprocessing techniques and the experiments' procedures. The results of the experiments are presented and thoroughly discussed in Section 5, and the discussion also includes comparisons of the proposed model with some existing standard methods and visualisation of the model output. Section 6 is the conclusion of this work.

2 | RELATED WORKS

Related Works The three main aspects of Facial Expression Recognition (FER) research are the data, the model and FER data annotations. Researchers have thoroughly considered diverse ways to resolve the limitation in FER from data and model points of view.

Initially, static data in their hundreds were the only available data in the field, and they were collected in a controlled environment. The laboratory collected data limits the performance of FER because FER developed in a controlled environment fails to be generalised to the real world. The introduction of sequence and temporal data give room for emotion recognition and the corresponding intensity estimation. The advent of deep learning, especially Convolution Neural Network (CNN), poses a great demand on the volume of learning data to make an appropriate FER model. With internet facilities, data collection is now accessible, which mitigates the challenges with data availability. Likewise, different models have been proposed in the literature for optimal FER. Among the existing models, deep learning models have records of superior performance. Variants of deep learning models consider for FER implementation include CNN, Generative Adversarial Network (GAN), Multitasking Convolution Neural Network (MCNN), Cascaded Convolutional Neural Network (CCNN), to mention a few. The deep networks have successfully contributed immensely to achieving a robust FER. Still, the deep networks only considered the logical labels in their classification task. Most of the existing methods fail to incorporate correlation among data labels. Using logical labels make the models not capture the ambiguous nature in FER datasets.

Recently, Multilabel Learning (ML) methods have been adopted in FER to account for FER label ambiguity. Group Lasso Regularised Maximum Margin (GLMM) proposed by Zhao et al. (2013), considered the fact that the Action Unit (AU) at different affective states is triggered in the same region of the face. GLMM used the feature extracted for different expressions at the same region to classify them into a zero or non-zero, making it possible for a group to contain different expressions. The global solution of the model was achieved by a function called Maximum Margin Hinge loss. GLMM was later enhanced to Adaptive Group Lasso Regression Zhao et al. (2014) to assign a continuous value to the distribution of expression present in a non-zero group. GLMM shows its superior performance compares with some existing ML methods from the experiment conducted on s-JAFFE. Li and Deng (2018) is another prominent multilabel approach to FER, LI and Deng Li and Deng (2018) introduced a multilabel deep learning model termed Deep Bi-Manifold CNN (DBM-CMM). The model preserves the local affinity of deep emotion features and the manifold structure of emotion labels, while learning the discriminating feature of multilabel expression. The deep network training is supervised by softmax cross-entropy loss jointly with the bi-manifold loss for feature discriminating enhancement. This model resolves data ambiguity accurately from RAF-ML data and generalised well with existing multilabel data through the incorporated adaptive mechanism. Nevertheless, the multilabel approach could only resolve the ambiguity problem but could not account for the intensity of the recognised emotions.

Label inconsistency and ambiguity challenges are lately considered using the LDL approach. LDL application to FER requires the construction of distributed labels of each of the facial expression instances; as found in Zhou et al. (2015) Xing et al. (2016) Xi et al. (2020). Jia et al. (Jia et al., 2019) preserved the correlation among FER data label locally using EDL-LRL (Emotion Distribution Label-Low Ranking label correlation Locally), which forms a low-rank structure that alleviates the complexity in emotion correlation, with an assumption that low-rank structure represents the label space. The experiments conducted on label distribution datasets (s-JAFFE and s-BU3DFE) show the proposed model's prominence. The model considers the correlation among the label locally on data with a distribution label. Abeere et al. Almowallad and Sanchez (2020) proposed a feature hybrid based model called EDL-LBCNN, which hybridised Local Binary Convolution (LBC) features and Convolution Neural Network (CNN) features train with Kullback-Leibler loss and optimise with ADMM (Alternating Direction Method of Multipliers). The outcome of the experiment on the s-JAFFE dataset shows its promising performance. Zhang et al. (Zhang et al., 2020) proposed a Correlated Emotion Label Distribution Learning (CELDL) model for infrared facial expression recognition. The model initially computes the correlation between expression images using cosine similarities and finally learns the basic emotion in infrared expression with deep CNN. This method is domain-specific, as the model performance was reported only on infrared features. The above LDL models require FER data with distribution labels, and the generalisation of the methods to in-the-wild data and data with a logical label is a challenge.

Chen et al. (Chen et al., 2020) proposed a label enhancement model to recover label distribution from the logical data annotations. The model generates an auxiliary label space from action units and facial landmarks. This method minimises the problem encountered in Graph Laplacian Label Enhancement (GLLE) model using approximate KNN for building the approximate KNN (akNN) graphs that generate the auxiliary labels. Deep CNN was used as the backbone of the proposed system. The experiments conducted on some laboratory-controlled data (CK+, Oulu-CASIA, CFEE, MMI) and in-the-wild (AFFNET, RAF, SFEW) data proved the system's efficiency over existing methods with an assurance of label consistency and consideration of label ambiguity. The model did not report the intensity of the recognised emotions. There is a need for methods in the field that could recover the distribution label from logical annotations of data to resolve label ambiguity, inconsistency and intensity estimation.

Graph Neural Networks (GNN) are presently gaining attention because of their efficiency in graphical structure problems. A thorough investigation of GNN models is available in Zhou et al. (2020). GCN proposed by Kipf and

Welling (2017) is one of the variants of GNN, and it has been used diversely in image-related tasks such as image classification Garcia and Bruna (2017), image semantic segmentation Liang et al. (2017), region classification Chen et al. (2018) and object detection Hu et al. (2017). GCN could be implemented for supervised, unsupervised or transductive learning depending on the task, the availability, and the integrity of the available data.

In this work, we propose a manifold learning model as a graph-based label enhancement approach to learn the correlation among data labels, and GCN is employed as a semi-supervised model to recover the emotion distribution from the logical labels of the datasets to address annotation inconsistency, ambiguity and expression intensity estimation challenges in FER.

3 | MANIFOLD GRAPH CONVOLUTIONAL NETWORK MODEL

This section contains manifold learning and graph convolutional network model description for facial expression recognition task.

3.1 | Isomap Manifold

Isomap manifold represents a high dimensional dataset in a low dimensional space with the preservation of the fundamental relationship among the data Tenenbaum et al. (2001) Samko et al. (2006). Facial expression data are non-linear and high dimensional, which make them candidate of dimensionality reduction algorithm. Assuming FER data X$\in$ $R^{M\times N}$, such that X = {x_1,x_2,........,x_n}. Isomap manifold tends to find a lower dimensional space m and embed X samples, that is, X$\in R^{m\times N}$. The manifold achieved the embedding of facial expression data into low dimensional space by using the global topological information of feature space to obtain similarity distance about data points x_i. Euclidean distance is reported to be susceptible to short circuit edge problem Chen et al. (2006), which implies that it could provide neighbours along the external space that are not neighbours along the manifold. To avoid the short circuit edge problem, and degradation in isomap performance in the presence of noise. We employ the image based euclidean distance proposed in Chen et al. (2006). Each expression image is first transformed linearly using (1) where σ the width parameter is taken to be 1. Assuming two images x and y, then the Euclidean distance between them is given as:

$$d^2(x,y) = \sum_{i,j=1}^{MN} g_{i,j}(x_i,y_i)^T G(x,y) \tag{1}$$

where the image vector g_{ij} = $\frac{1}{2\pi\sigma^2}\exp[\frac{-(|P_i-P_j|^2)}{2\sigma^2}]$ and symmetric matrix G= $(g_{ij})_{MN\times MN}$, P_i, P_j are pixels and $|P_i$-$P_j|$ is the distance P_i and P_j on the image lattice. Also u = $G^{\frac{1}{2}}$x and $G^{\frac{1}{2}}$ = $\alpha\,\gamma^{\frac{1}{2}}\,\alpha$. α is the orthogonal matrix whose column vector are eigenvectors of G and $\gamma^{\frac{1}{2}}$ is the diagonal matrix which contains the eigenvalue of G.

The distances between the transformed images are computed from (2).

$$d(x) = (x_i - x_j)^T G(x_i - x_j) = (u - v)^T (u - v) \tag{2}$$

The nearest neighbourhood graph is constructed from the computed distances d(x), and the Dijkstra algorithm (the shortest algorithm) finds the shortest path along the neighbourhood graph to compute non-neighbouring data points. The application of multidimensional scaling generates the low dimensional vector space.

3.2 | Graph Convolution Network

Graph Convolution Network (GCN) has intrinsic ability to work directly on a graph related problem, and this motivates us to apply GCN for node classification using the available information through Isomap manifold.

Generally, GCN works by propagating node features along edges of the neighbouring graph, through which it learns the latent representative features in each node for node classification. Assuming a data X with N×F matrix and N×N adjacency matrix G, where N is the number of X samples, and F is the number of features in each sample of X, then GCN latent layers are given as:

$$H^i = f(H^{(i-1)}, A) \tag{3}$$

Where f is GCN propagation rule which computes a node features from the neighbouring nodes without including the node information and transform it with the application of weight W and the activation function σ, the expression is shown in (4) below:

$$f(H^i, A) = \sigma(AH^{i-1}W^{i-1}) \tag{4}$$

To include the node feature in the node aggregation in (2), A is modified by adding Identity matrix to A as $\hat{A}$ = A + I, and also multiply with the inverse degree of $\hat{A}$ denoted as D^{-1}, D^{-1} in (5) to resolve the possible problem attributed to gradient explosion, which could emanate from the degree of the node because the feature value increases with the node degree. Multiplying D^{-1} with the representing features normalises the feature representation, and thus making the feature representation aggregation invariant to node degree.

$$f(H^i, A) = \sigma(D^{-1}\hat{A}D^{-1}H^{i-1}W^{i-1}) \tag{5}$$

The adjacency matrix only assigns one as the edge weight in the original GCN, which would possibly misguide the label's prediction in the presence of adjacent nodes of different labels. To avoid this, a similarity graph that assigns different edge weights to different neighbour nodes should be used with GCN as an aggregate of neighbour nodes' information.

This work uses distance generates along the Isomap manifold as the similarity graph for GCN aggregation of neighbour nodes' information. The objective function is computed using a multidimensional scale that runs on the geodesic distance to generate the co-ordinate of the Euclidean space. Given the Multidimensional Scaling (MDS) function z $\in R^{n\times k}$ then the objective function in (7) is obtained from the derivative of (6).

$$MDS = argmin\frac{1}{2}\sum_{n}^{i=1}\sum_{n}^{i=i+1}(||x_i - x_j|| - ||z_i - z_j||)^2 \tag{6}$$

where $||x_i\text{-}x_j||$ is the original distance in d-dimensional space, and $||z_i\text{-}z_j||$ is the k-dimensional data points such that the original distances are preserved.

$$f(z) = \sum_{n}^{i=1} \sum_{n}^{i=i+1} s_{ij}\Big(-\frac{z_i - z_j}{\sqrt{(z_i - z_j)^2}}\Big) \tag{7}$$

where s_{ij} = || x_i-x_j|| - || z_i-z_j|| = $\sqrt{(x_i - x_j)}^2$ -$\sqrt{(z_i - z_j)}^2$. The Isomap distances at the lower dimensional space also preserve distance and similarity proportionality relationships; that is, the higher the distance, the lower the similarity between nodes and the lower the distance, the higher the similarity. The computed similarity matrix represents the edge weight between nodes, which in turn help to construct a graph with the appropriate information about adjacent nodes.

If the computed similarity graph is represented with Θ=Θ_{ij}, where Θ_{ij} is the similarity between $node_i$ and $node_j$. Then, replacing the adjacency graph A in (4) with the computed similarity graph, Θ generate an updated version of (4) as presented in (8).

$$f(H^i, \Theta) = \sigma(\Theta H^{i-1} W^{i-1}) \tag{8}$$

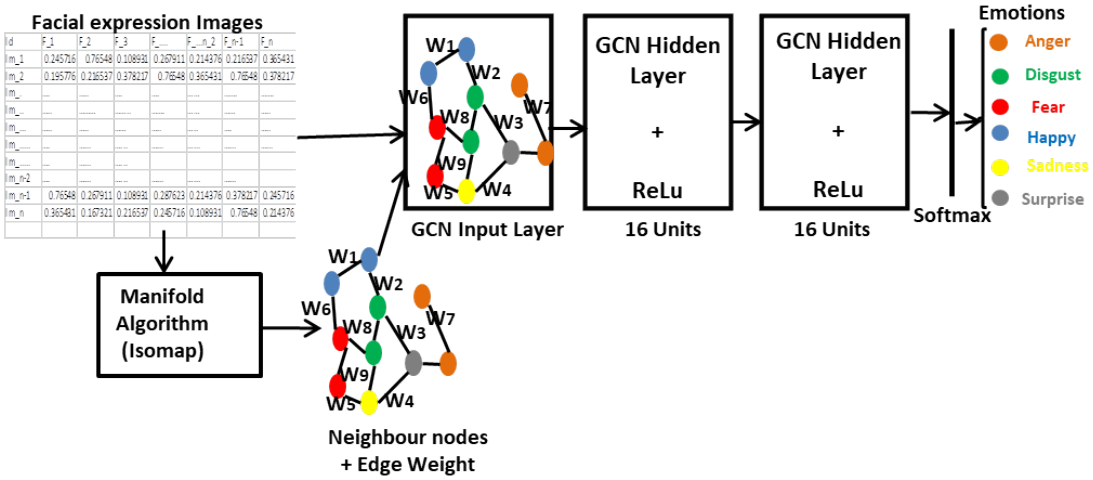
The similarity graph models the correlations among emotions, assuming that data points close in the low dimensional space should have the same label description. Given the feature matrix E and the manifold similarity graph θ, GCN could predict the required emotion distribution in facial expression images. The output layer of GCN is now expressed in (9), which is the final output weight given as W^i

$$W^i = softmax(\sigma(\Theta E W^{i-1})) \tag{9}$$

Algorithm 1: Manifold-GCN Algorithm

```
Input: : Expression images X ∈ $R^n$, k (number of Kneighbour nodes)
Output: : description degree of y to x such that $\sum_y d^y = 1$
Compute the linear transformation of each expression image $x_i$ ∈ X by multiplying $x_i$ with the eigenvalue
  and eigenvector of the symmetric matrix of the image vector.
Compute distances for every pair of $x_i$, $x_j$ using equation (2).
Generate neighbouring graph with geodesic algorithm.
Find the shortest path for every none neighbouring data points using Dijkstra algorithm.
the similarity graph $\theta = \theta_{i,j} = d_G(x_i, x_j)$.
Replace Adjacency graph ∈ f($H^i$, A) with the similarity graph $\theta$
Perform the classification with softmax function as given in equation (9)
```

Fig. 1 is the model framework description.



Id	F_1	F_2	F_3	F_....	F_...n_2	F_n-1	F_n
Im_1	0.245716	0.76548	0.108931	0.267911	0.214376	0.216537	0.365431
Im_2	0.195776	0.216537	0.378217	0.76548	0.365431	0.76548	0.378217
Im_.	...	...	...	...	...	...	...
Im_..	...	...	...	...	...	...	...
Im_...	...	...	...	...	...	...	...
Im_.....	...	...	...	...	...	...	...
Im_.....	...	...	...				
Im_n-2	...	...	...	...	...	...	
Im_n-1	0.76548	0.267911	0.108931	0.287623	0.214376	0.378217	0.245716
Im_n	0.365431	0.167321	0.216537	0.245716	0.108931	0.76548	0.214376

FIGURE 1 Architectural description of manifold GCN model.

4 | EXPERIMENT

This section presents details of the experiments from data preprocessing, feature representation and label enhancement and emotion classification.

4.1 | Datasets

This study considers two FER datasets, which are BU-3DFE Yin et al. (2006) and Cohn Kanade Extension (CK+) Lucey et al. (2010).

Binghampton University -3D Facial Expression (BU-3DFE)

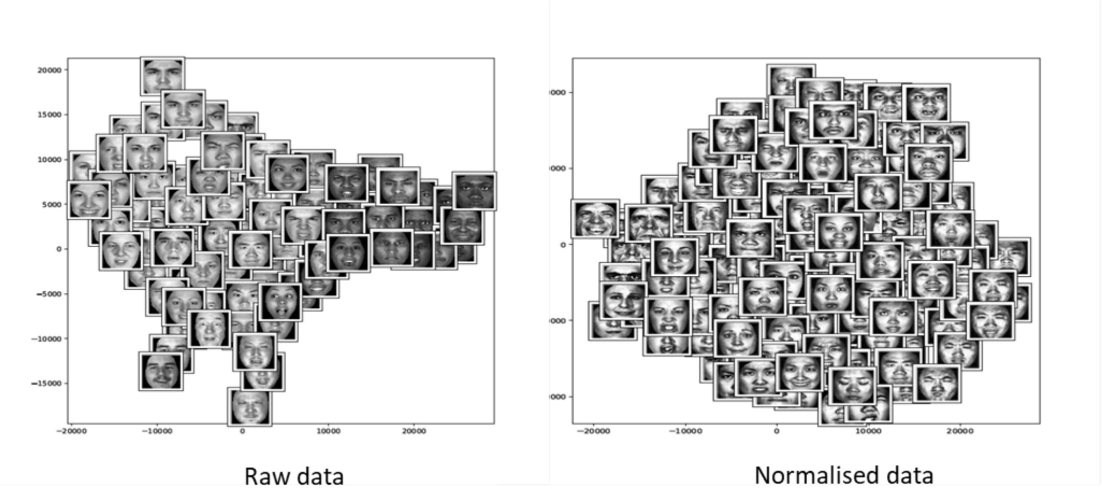
contains 2500 emotion data collected from 100 subjects in which 40% are male, and the remaining 60% are female. Although BU-3DFE is static and controlled, it is regarded as in-the-wild-data because of the different races, ethnicities, and ages of subjects in the data.

Cohn Kanade (CK+)

CK+, unlike the BU-3DFE dataset, is a sequence dataset; it consists of 7 expressions (Anger, Disgust, Contempt, Fear, Happy, Neutral, Sadness, Surprise) and the neutral face. CK+ contains 593 frames, whereby 327 are labelled out of them. The frames are produced by 127 subjects, of which 65% are female, and 35% are male. CK+ has been employed severally, virtually in all categories of facial expression recognition tasks. In this study, we are only considering six basic emotions.

4.2 | Data Preprocessing

Isomap performance is affected by noise. To ameliorate this effect, we employed a face detection algorithm and normalisation techniques available in OpenCV Bradski (2000) to minimise redundant information in the data. The Haar-like cascades are used for face detection, and a standard equalisation algorithm is used for contract normalisation.



Raw data

Normalised data

FIGURE 2 Figure showing Isomap transformation performance of raw expression images on the left and the preprocessed images of BU-3DFE on the right.

TABLE 1 The summary of data distribution in the Manifold GCN model.

DATA	Nodes	Edges	Classes	Train	Evaluation	Test
BU-3DFE	2400	2400	6	50	150	2200
CK+	593	593	6	50	100	443

Fig.2 shows the information preserved in BU-3DFE data before and after preprocessing. Figure 2a shows the effect of light intensity variation on the manifold visualisation of unpreprocessed data, and Figure 2b reflect the importance of the application of light normalization to the data in manifold visualisation.

4.3 | Experimental Setup

The images are the graph's nodes; the edges and their respective weight are generated via the Isomap manifold. We achieved the model implementation by using stellar graph Data61 (2018) and keras with tensorflow 2.2 as backend Chollet et al. (2017). ADAM optimiser Kingma and Ba (2014) is employed at the training phase of the model with a learning rate of 0.03, dropout of 0.25, weight decay (L2 regularisation) of $5e^{-4}$. The Manifold GCN model has two hidden graph convolutional layers with ReLu activation function, and the output layer for nodes classification is the softmax activation function. For each of the datasets, less than 10% are randomly selected as label data, 16% as evaluation data, and the remaining data are used as test data. The summary of the data is available in Table 1. The training process takes a maximum of 100 epochs, and we also use the early stopping function to stop the training after a consistent increase in loss values. The experiment results are presented from an average of 10 runs.

The first phase of the experiments is the model performance evaluation with parameter tuning, and we observe the system's performance with different K values, optimisers, and learning rates. Furthermore, we compare the system's performance with some existing models developed for label enhancement distribution learning for FER. We conclude the experiment by visualising the embedding vectors in our model.

TABLE 2 The Performance of the manifold GCN model on datasets based on test accuracy(%) and loss values.

Model	Database	Accuracy	F1_Score	Average Precision	Average Recall
Manifold GCN	BU-3DFE	98.80%	0.97	0.97	0.98
	CK+	97.20%	0.96	0.96	0.97

5 | EXPERIMENTAL RESULTS AND DISCUSSION

The training and the evaluation accuracy and loss of the model are shown in Fig. 3.The model performance on the test datasets in this work is presented as multiclass confusion matrices in Fig. 4 and Fig. 5, and from the confusion matrices we compute accuracy, F1-Score, average precision, average recall metrics as given in equation 10, equation 11, equation 12 and equation 13 respectively. The outcomes of the computations are detailed in Table 2. The comparison study with some existing models is presented in Table 3, and we compare the model with some baseline methods on the datasets using accuracy as the metrics for performance evaluation. The proposed method has an outstanding performance on BU-3DFE data. The model performance on CK+ is promising compare to other models. This implies that consideration of similarity measures accounted for by our model positively affects the system's performance.

$$accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{10}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{precision + Recall} \tag{11}$$

where Precision = $\frac{TP}{TP+FP}$ and recall = $\frac{TP}{TP+FN}$.

$$AveragePrecision = \frac{TotalPrecision}{no_of_classes} \tag{12}$$

$$averagerecall = \frac{Totalrecall}{no_of_classes} \tag{13}$$

TP simply means True Positive, TN means True Negative, FP means False Positive and FN means False Negative.

Experiment result for k values and parameter setting

We consider increasing the number of k(neighbour) parameters and observe the effect on the model performance. K values are increased to the multiple of two, that is, set {2,4,6,8} as presented in Table 4. We observe that the model gives stable and optimal results for the datasets when k= 4 and deduce that the more the neighbour size, the more difficult the classification task. This is suspected to be a result of accommodating members with different features that introduce information that misguides the central node classification. For our model parameters, we conducted a manual search between three optimisation functions (Stochastic Gradient Descent (SGD), ADADELTA, ADAM) and learning rate was varied between 0.1 and 0.5 step by 0.1, 0.001 and 0.005 step by 0.001, and lastly 0.0001 and

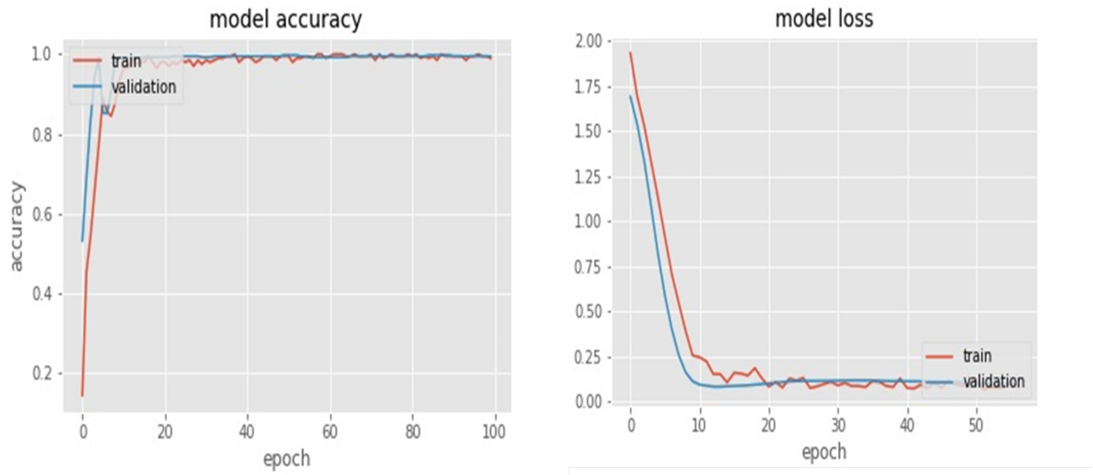


FIGURE 3 The right figure is the training and validation accuracy graph, while the left figure is the training loss and the validation loss of the model performance.

TABLE 3 The Performance of the manifold GCN model compare with existing models on CK+ data.

Author	Method and contribution	Number of classes	Accuracy
Wu et al. (2017)	E3DNET	7	80.36
Li and Deng (2018)	DBMNET	7	85.27
Ekundayo and Viriri (2019)	Deep Forest	6	93.07
Chen et al. (2020)	LDL-ALSG	7	93.08
Liu et al. (2021)	CDLLNET	7	87.42
Gan et al. (2020)	MA-FER	7	83.28
Proposed Model	**Manifold GCN**	**6**	**97.20**

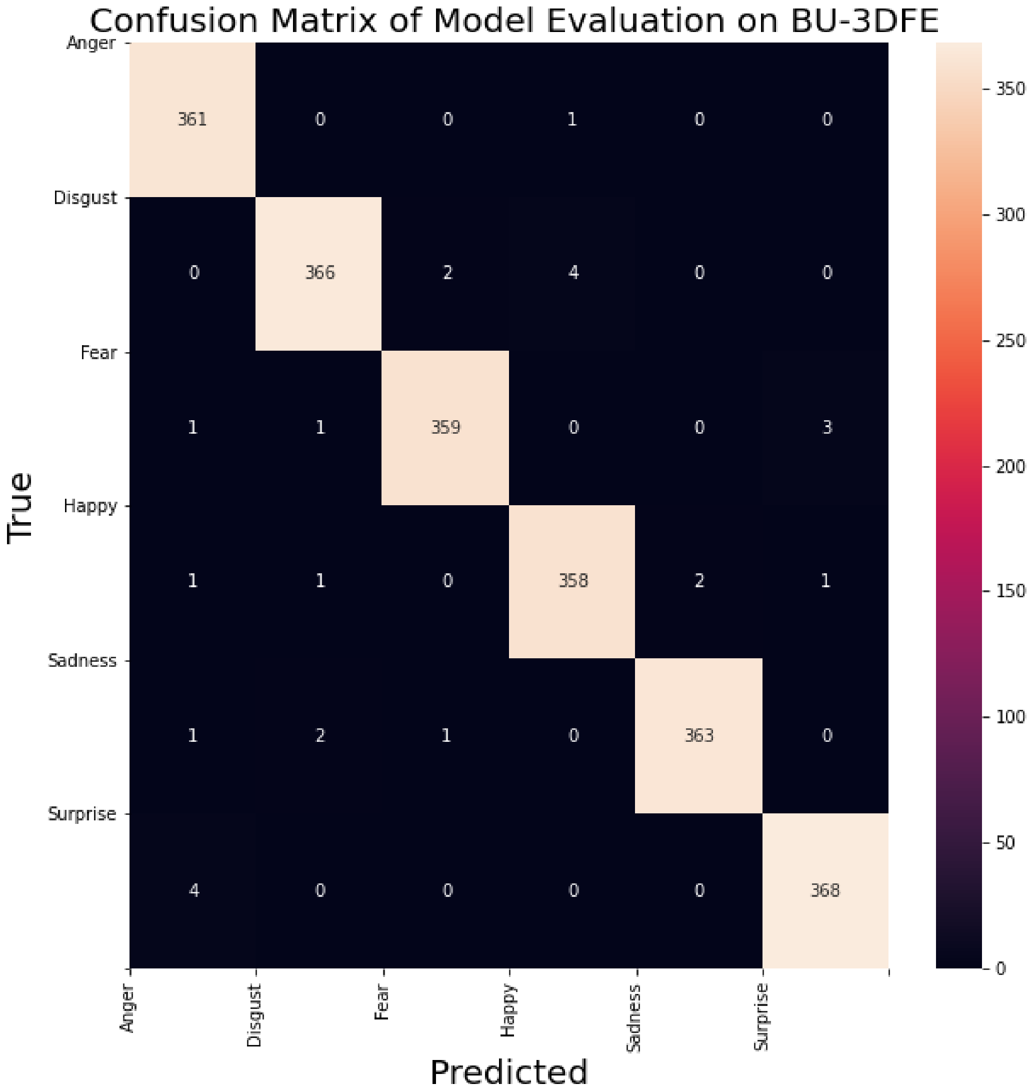


FIGURE 4 Confusion matrix of the prediction of MGCN model on BU-3DFE test data

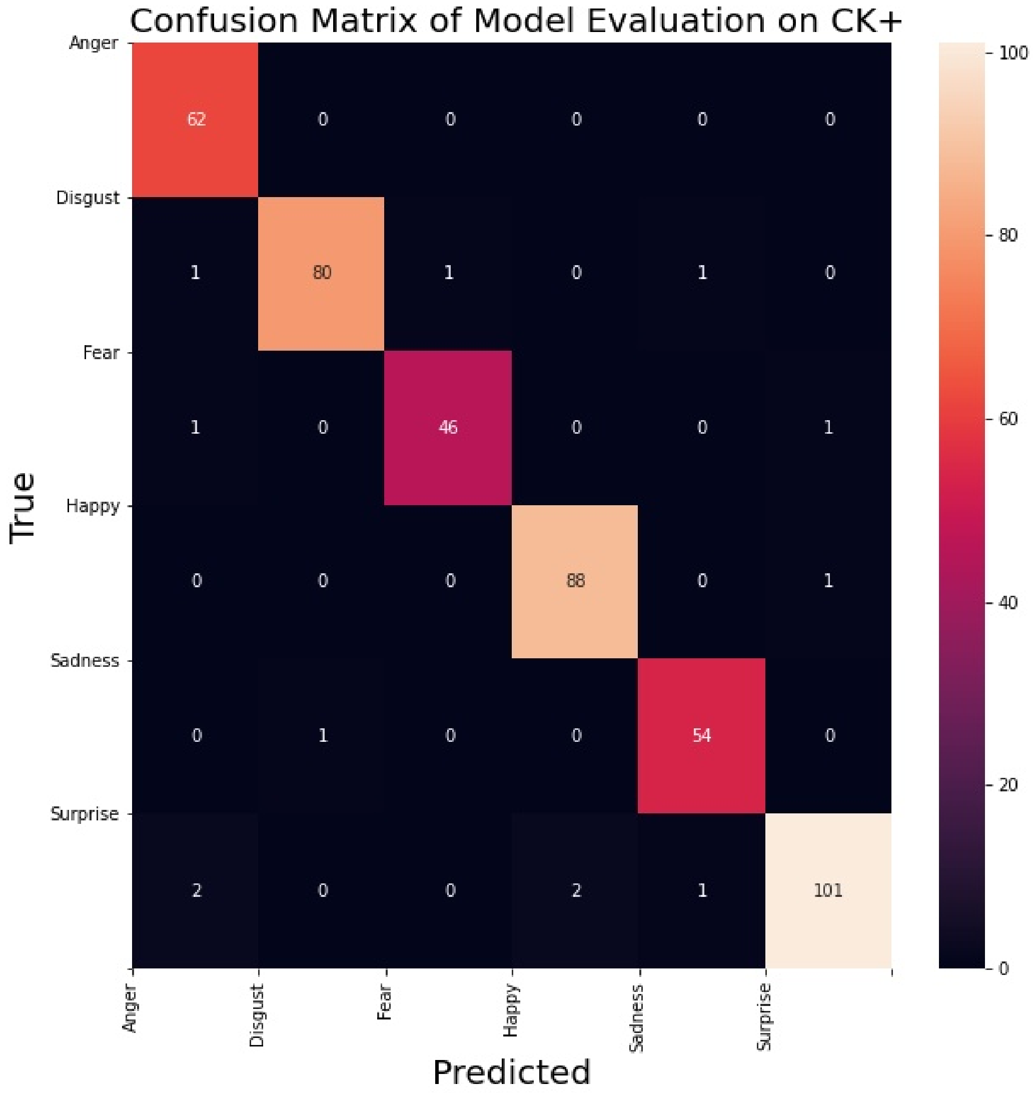


FIGURE 5 Confusion matrix of the prediction of MGCN model on CK+ test data

TABLE 4 Table showing the model performance at different values of k.

DATABASE	K = 2	K = 4	K = 6	K = 8
BU-3DFE	94.03%	98.80%	87.53%	80.16%
CK+	94.72%	97.20%	86%	72.59%

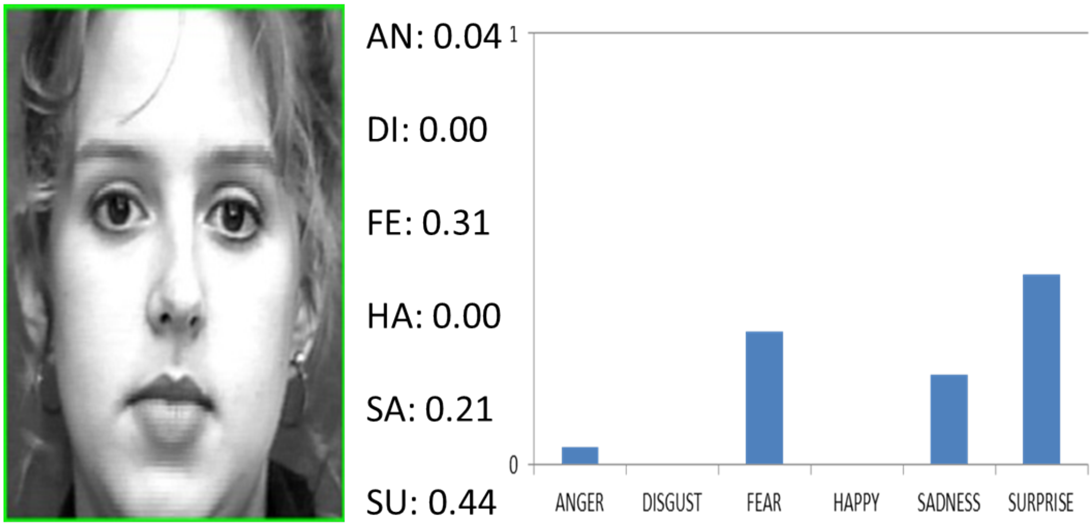


FIGURE 6 Sample of manifold GCN prediction of CK+ data

0.0005 step by 0.0001. We also search for the dropout value between 0.25 and 0.5. We obtained optimal results with ADAM optimiser, 0.003 learning rate, and 0.25 dropout.

Model Visualization

Graph convolutional network is implemented as a semi-supervised model, couple with manifold enhancement the model is capable of learning correlation among labels and hence resolves data inconsistencies. The model's output is the softmax activation function, softmax approximately gives the proportion of each emotion in the expression face, and removing the threshold enables the model to recover the emotion distribution from logical labels. A sample of the model prediction is presented in Fig. 6. We employed Isomap low dimensional data embedding technique to visualise features in the hidden layers of our model. Isomap has the tendency to map high dimensional data feature with similar label to the corresponding low dimensional space. The classification output is viewed based on the categories of different colours, where each colour represents a particular class. Fig 7 provides information about the hidden layers of the proposed model. The closeness of colours with different labels speaks volume about the correlation and the distribution of emotion in facial expression.

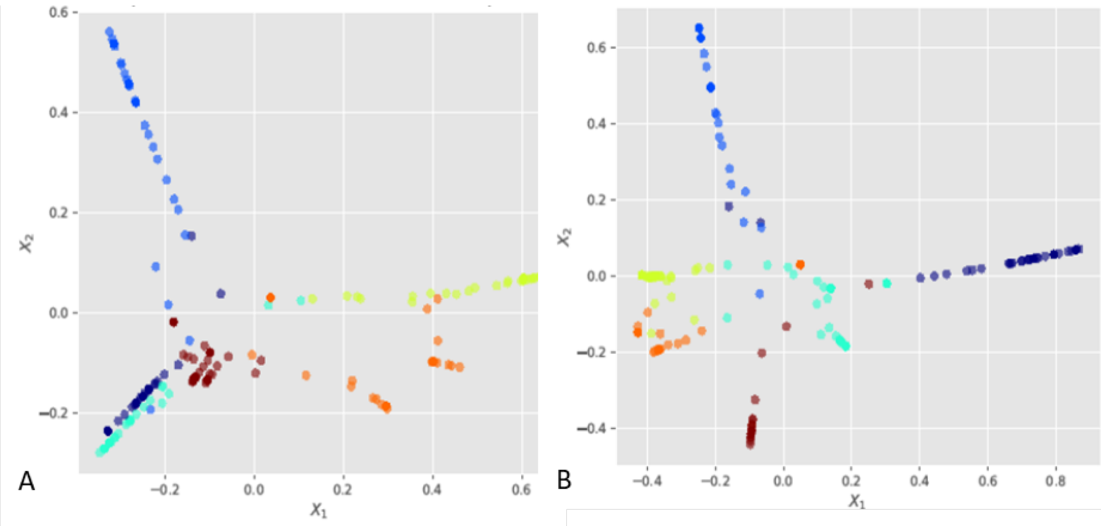


FIGURE 7 A is the 2-dimensional visualization result of the model information of BU-3DFE data, and B is the 2-dimensional visualization of what the model learns of CK+ data.

6 | CONCLUSION

This work implements label enhancement for facial expression distribution learning. We adopt inductive technique, which is manifold learning to account for the similarities and correlations among FER data, which help compute distance similarity for graph convolution networks to substitute for adjacency matrix. The model performs optimally on both the CK+ and BU-3DFE datasets, visualising the data with Isomap dimensional reduction indicates that the model learns the correlation appropriately among data classes and hence resolve data inconsistencies. This study is conducted in a controlled environment, that is, on laboratory prepared data. Future work would consider more challenging environments like data in the wild or dynamic environments. The computation of distance similarities is an inductive process in isolation from the semi-supervised learning GCN, such that GCN could only make predictions based on the prior knowledge of seen data.

conflict of interest

Authors claim that there is no conflict of interest.

references

Almowallad, A. and Sanchez, V. (2020) Human emotion distribution learning from face images using cnn and lbc features. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 1–6.

Bradski, G. (2000) The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chen, J., Wang, R., Shan, S., Chen, X. and Gao, W. (2006) Isomap based on the image euclidean distance. vol. 2, 1110–1113.

Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X. and Rui, Y. (2020) Label distribution learning on auxiliary label space graphs for facial expression recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13981–13990.

Chen, S.-B., Tian, X.-Z., Ding, C., Bin, L., Liu, Y., Huang, H. and Li, Q. (2020) Graph convolutional network based on manifold similarity learning. *Cognitive Computation*, **12**, 1144–1153.

Chen, X., Li, L.-J., Fei-Fei, L. and Mulam, H. (2018) Iterative visual reasoning beyond convolutions. 7239–7248.

Chollet, F., Allaire, J. et al. (2017) R interface to keras. `https://github.com/rstudio/keras`.

Data61, C. (2018) Stellargraph machine learning library. `https://github.com/stellargraph/stellargraph`.

Ekman, P. and Friesen, W. V. (1971) Constant across cultures in the face and emotion. **17**, 124–129.

Ekundayo, O. and Viriri, S. (2019) Deep forest approach for facial expression recognition. In *PSIVT 2019 International Workshops*, 149–160. Sydney, NSW, Australia,.

Gan, Y., Chen, J., Yang, Z. and Xu, L. (2020) Multiple attention network for facial expression recognition. *IEEE Access*, **8**, 7383–7393.

Garcia, V. and Bruna, J. (2017) Few-shot learning with graph neural networks.

He, T. and Jin, X. (2019) Image emotion distribution learning with graph convolutional networks. 382–390.

Hu, H., Gu, J., Zhang, Z., Dai, J. and Wei, Y. (2017) Relation networks for object detection.

Jia, X., Zheng, X., Li, W., Zhang, C. and Li, Z. (2019) Facial emotion distribution learning by exploiting low-rank label correlations locally. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9833–9842.

Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. *CoRR*, **abs/1412.6980**. URL: `http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14`.

Kipf, T. N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17. URL: `https://openreview.net/forum?id=SJU4ayYgl`.

Li, S. and Deng, W. (2018) Blended emotion in-the-wild : multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, **127**, 884–906. URL: `https://doi.org/10.1007/s11263-018-1131-1`.

Liang, X., Lin, L., Shen, X., Feng, J., Yan, S. and Xing, E. (2017) Interpretable structure-evolving lstm. 2175–2184.

Liu, T., Wang, J., Yang, B. and Wang, X. (2021) Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. *Infrared Physics and Technology*, **112**, 103594.

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. 94 – 101.

Samko, O., Marshall, D. and Rosin, P. (2006) Selection of the optimal parameter value for the isomap algorithm. *Pattern Recognition Letters*, **27**, 968–979.

Tenenbaum, J., Silva, V. and Langford, J. (2001) A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, **290**, 2319–23.

Wu, Z., Chen, T., Chen, Y., Zhang, Z. and Liu, G. (2017) Nirexpnet: Three-stream 3d convolutional neural network for near infrared facial expression recognition. *Applied Sciences*, **7**. URL: `https://www.mdpi.com/2076-3417/7/11/1184`.

Xi, X., Zhang, Y., Hua, X., Miran, S. M., Zhao, Y.-B. and Luo, Z. (2020) Facial expression distribution prediction based on surface electromyography. *Expert Systems with Applications*, **161**, 113683. URL: `http://www.sciencedirect.com/science/article/pii/S0957417420305078`.

Xie, G., Shi, H., Yin, B., Kang, Y., Shao, C. and Gui, J. (2017) Robust l-isomap with a novel landmark selection method. *Mathematical Problems in Engineering.*, **2017**.

Xing, C., Geng, X. and Xue, H. (2016) Logistic boosting regression for label distribution learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4489–4497.

Yin, L., Wei, X., Sun, Y., Wang, J. and Rosato, M. (2006) A 3d facial expression database for facial behavior research. vol. 2006, 211– 216.

Zhang, Z., Lai, C., Liu, H. and Li, Y.-F. (2020) Infrared facial expression recognition via gaussian-based label distribution learning in the dark illumination environment for human emotion detection. *Neurocomputing*, **409**, 341 – 350.

Zhao, K., Zhang, H., Dong, M., Guo, J., Qi, Y. and Song, Y.-z. (2013) A multi-labelclassification aprroach for facial expression recognition. In *Visual Communications and Image Processing*. Kuching, Malaysia: IEEE.

Zhao, K., Zhang, H. and Guo, J. (2014) An adaptive group lasso based multi-label regression approach for facial rxpression analysis. In *International Conference on Image Processing(ICIP)*, 1435–1439. Paris, France: IEEE.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M. (2020) Graph neural networks: A review of methods and applications. *AI Open*, **1**, 57–81. URL: `https://www.sciencedirect.com/science/article/pii/S2666651021000012`.

Zhou, Y., Xue, H. and Geng, X. (2015) Emotion distribution recognition from facial expressions. In *MM ’15: Proceedings of the 23rd ACM international conference on Multimedia*, 1247–1250.

3.5.2 Conclusion

The proposed Manifold GCN performance was encouraging for the facial expression datasets (CK+ and BU-3DFE). This showed that learning the correlation among labels to approach data annotation inconsistency and capturing facial expression label ambiguity is influential for an optimal facial expression recognition system.

Chapter 4

Results Presentation and Discussions

4.1 Introduction

This chapter presents the outcomes of the papers presented in Chapter 3. Firstly, the Deep Forest model results and the performance comparison with the existing models for emotion classification is presented. Next is the Multilabel Convolutional Neural Network model's results for emotion recognition and ordinal intensity estimation. Finally, the Manifold Graph Convolutional Network experimental results are presented and their comparison with some existing models. The experiments were conducted on BU-3DFE and CK+ datasets. The hardware specification for deep forest computation include: (intel(R)Core(TM)i7-4770sCPU @3.10 GHz 3.10GHz and RAM: 8GB) Dell machine. CHPC high computing device was used for Multilabel Convolution neural network and Manifold graph convolution network was computed on (intel(R)Core(TM)i7-4770sCPU @3.10 GHz 3.10GHz and RAM: 8GB) Dell machine.

4.2 Deep Forest Approach for Facial Expression Recognition

The Deep Forest model was used for facial expression classification to achieve deep learning with shallow classifiers to accommodate the small data size in the field. The method employs different forest tree algorithms to learn in layers until there is no significant performance increase in the output of successive layers. The two datasets considered for the research experiments are the BU-3DFE and CK+ data. Only the peak expressions for each emotion class were employed for the two datasets, while the model performance evaluation metrics were confusion matrix and accuracy, Figure 4.1 and Figure 4.2 provide the details of the experimental results. Furthermore, Deep Forest performance was also compared with some existing models that considered the datasets in their experiments, this is available in Table 4.1.

Table 4.1: the result comparison of FERAtt (Facial Expression Recognition with Attention Net) with Deep Forest learning

Author	Database	Accuracy
Fernandez et al.[36]	BU-3DFE	75.22%
hariri et al [37]	BU-3DFE	92.62
Derkach and Sukno [38]	BU-3DFE	81.5%
Yin et al.[1]	BU-3DFE	83%
Our	BU-3DFE	65.53%
Saeed et al. [39]	CK+	83.01
Fernandez et al.[36]	CK+	86.67%
Uddin et al. [40]	CK+	93.23%
Our	CK+	93.22%

Figure 4.1 contains the graphs of the average recognition rate on the test data of BU-3DFE and CK+. Also, Figure 4.2contains the confusion matrices of the model probabilistic predictions accuracy on the CK+ and the BU-3DFE, respectively.

Figure 4.2 BU-3DFE confusion matrix shows that the prediction of the model is most for the surprise at 95%. Followed by happy at 90% then disgust at 55%, both sad and fear have 50% prediction accuracy and angry has the least prediction at 40%. While CK+ shows that the model gives 100% prediction for angry, disgust, fear and happy instances, 94% for surprise and 40% for sad.

The performance of Deep Forest was justified based on facial expression classification by comparing its performance with the state-of-the-art DNN method(FERAtt) [36]. Table 4.1 presents both the Deep Forest results and FERAtt results. Deep Forest gives better accuracy (93.22%) than the accuracy achieved in FERAtt (86.67%) on the CK+ dataset. While accuracy gotten with FERAtt(75.22%) on the BU-3DFE dataset is more than Deep Forest (65.53%). However, it should be noted that FERAtt could not use a small dataset because the authors reported that the data were augmented and also combined with Coco data. Similarly, FERAtt demands high computing device like GPU for its appreciable time of computation, unlike the Deep Forest that performed its layer by layer learning on the available computing device (intel(R)Core(TM)i7-4770sCPU @3.10 GHz 3.10GHz and RAM: 8GB) at a reasonable time of computation.

4.3 Multilabel Convolutional Neural Network for Facial Expression Recognition

This section presents a multilabel-CNN model which detects the intensity estimation along with the emotion recognition concurrently. The model employed binary relevance multilabel approach at the first phase, where each expression image is associated with the respective ordinal intensity metrics. Convolutional Neural Net-

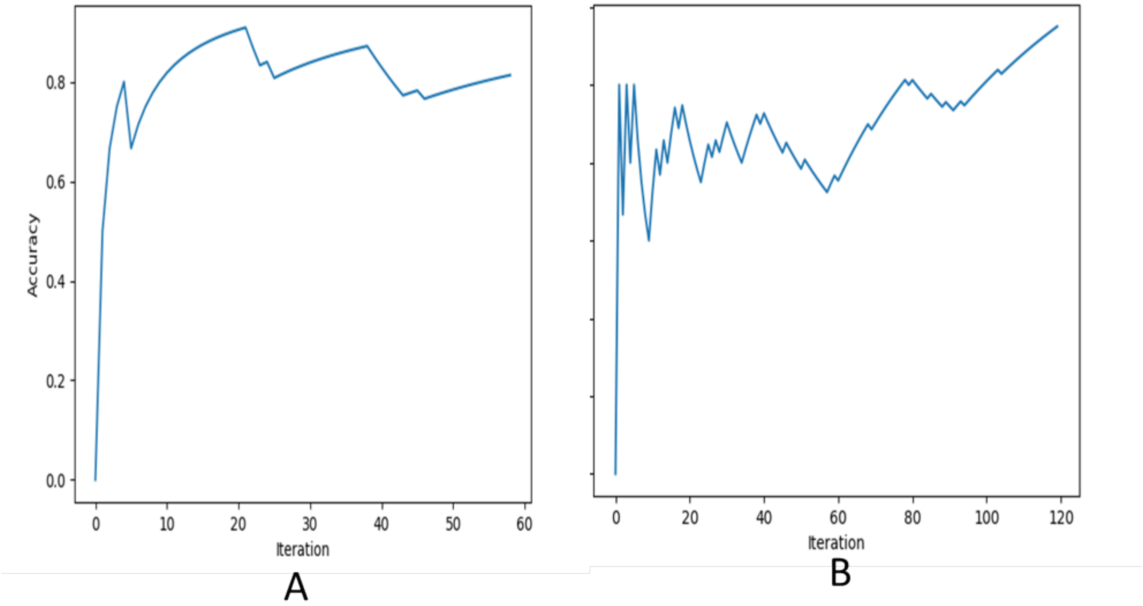
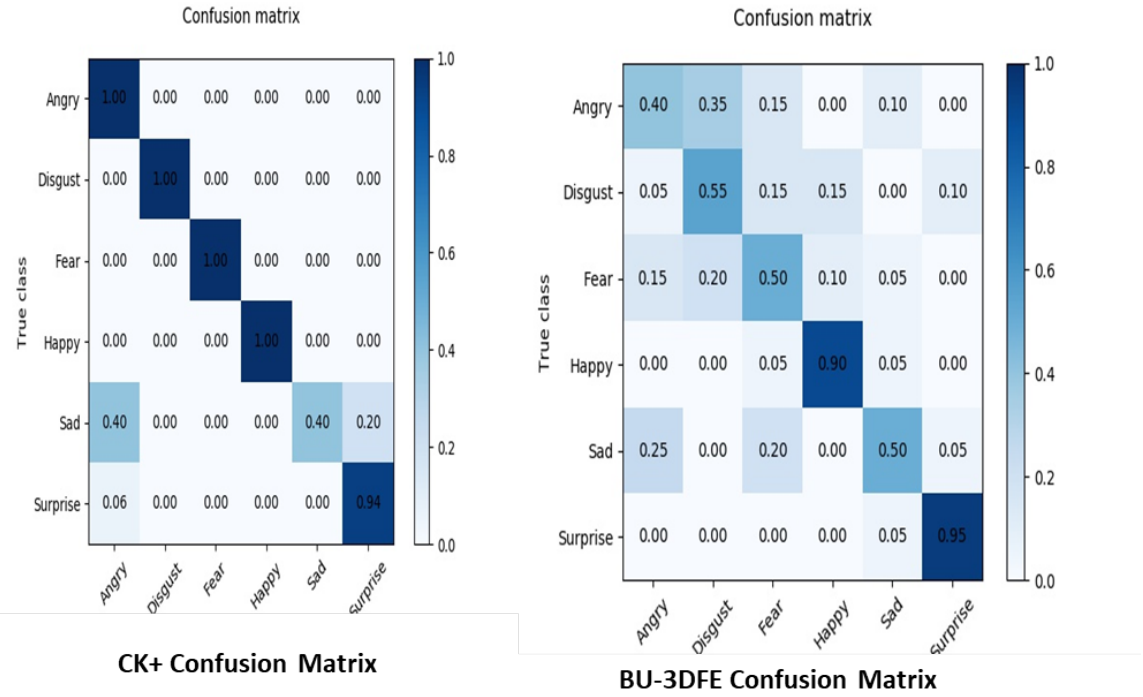


Figure 4.1: Deep Forest accuracy graph: A is the accuracy graph for CK+ and B is the accuracy graph for BU-3DFE.

work is employed at the second phase to learn the multilabel structure and predict the emotion and intensity based on ordinal metrics. The overfitting effect was minimised in the multilabel model presented in section 3.3 with the VGG-16 pre-trained network. The interclass and intraclass variations were also reduced with an island loss function. The output layer of the model is the sigmoid activation function, which presents the independent probability of the available emotion and their ordinal intensity. The results of the model performance on BU-3DFE and CK+ datasets are presented in this section. The proposed Multilabel-CNN performance evaluation is based on accuracy and loss function in Table 4.2 and Figure 4.3.

Multilabel-CNN is also compared with some other multilabel models: RAKELD (Distinct Random k-Label sets)[41], classifier chain (CC)[42], MLkNN (Multilabel k Nearest Neighbour) [43] and MLARAM [44]. Multilabel metrics used for the comparison include hamming loss, average precision, coverage error, and ranking loss. Details of the comparison are presented in Table 4.4 The model result was also compared with some recent deep learning networks for FER classification that consider the same datasets; details of the comparison is available in Table 4.5.

VGGML-CNN model predictions of some test data for BU-3DFE and CK+ are shown in Figure 4.3 and Figure 4.4, respectively. The label inscription on the image is the VGGML-CNN's predictions and the label tag at the base of the image is the actual label. The recognition rate of each emotion and ordinal intensities



CK+ Confusion Matrix

BU-3DFE Confusion Matrix

Figure 4.2: Deep Forest confusion Matrix for CK+ and BU-3DFE.

were finally presented using a multilabel confusion matrix shown in Figure 4.5 and Figure 4.6. The confusion matrices showed the prediction correctness of emotion with the associate intensity as they are visually presented in Figure 4.3 and Figure 4.4.

Table 4.2 presents the results of the base model (ML-CNN) and the optimised model with VGG-16 pretrained network VGGML-CNN. ML-CNN gives an average accuracy of 88.56% on BU-3DFE, 92.84 on augmented BU-3DFE and 93.24 on the CK+ dataset. At the same time, the VGGML-CNN increases the model performance on BU-3DFE by 6.3%, on augmented BU-3DFE by 5.6% and CK+ by 4.2%.

The results presented in Table 4.3 shows that the optimised multilabel CNN (VGGML-CNN) provides optimal results across the comparison metrics (hamming loss: 0.0890, ranking loss: 0.1647, average precision: 0.7093 and coverage error:1.9091). VGGML-CNN results for BU-3DFE are the bold values in the Table. Likewise for augmented BU-3DFE, VGGML-CNN outperformed all the multilabel algorithms considered, with hamming loss of 0.0628, raking loss of 0.1561, average precision of 0.8637 and coverage error of 1.3140. These are presented as bold values in Table 4.4. From Table 4.3 and Table 4.4, both the baseline model (ML-CNN) and the optimised version (VGGML-CNN) outperformed RAKELD, CC, MLkNN, and MLARAM. Table 4.5 and Table 4.6 show that VGGML-CNN outperformed most of the recent models' performances on CK+ and BU-3DFE with the accuracy of 97.16 and 98.01 respectively. Just like Xu et al. [45] and Chen et al. [34], The multilabel CNN provides the intensity estimation concurrently with the recognised emotion.

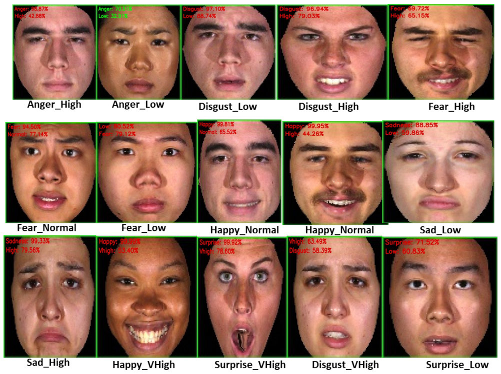


Figure 4.3: VGGML-CNN predictions of BU-3DFE [1].

4.4 Facial Expression Recognition with Manifold Learning and Graph Convolutional Network

This section presents the Manifold Graph Convolutional Network (Manifold-GCN) results and compares the results with the performance of some existing models. Manifold GCN models resolve label inconsistencies and ambiguity with the aid of manifold learning and GCN. Manifold learning uses similarity measures to model the correlation among data annotations and GCN learns the distribution of data labels by propagating data features together with the information of the neighbouring nodes. FER is transformed into a graphical problem where vertices are the data points and edges are the similarity distances. GCN model is made up of two hidden GCN layers with a softmax output layer. The performances of the model on BU-3DFE and CK+ are evaluated based on accuracy and ROC AUC. These results

Table 4.2: Presentation of ML-CNN and VGGML-CNN performance evaluation Using accuracy and aggregate loss on BU-3DFE and CK+ datasets

ML-Models	Database	Accuracy↑	Aggregate Loss ↓
ML-CNN	BU-3DFE	88.56	0.3534
	AUG_BU-3DFE	92.84	0.1841
	CK+	93.24	0.2513
VGGML-CNN	BU-3DFE	94.18	0.1723
	AUG_BU-3DFE	98.01	0.1411
	CK+	97.16	0.1842

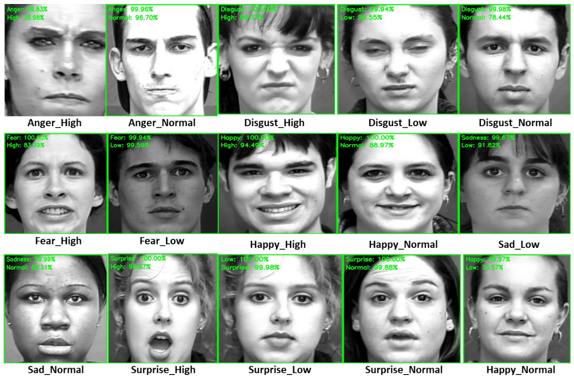


Figure 4.4: VGGML-CNN predictions of CK+ [2].

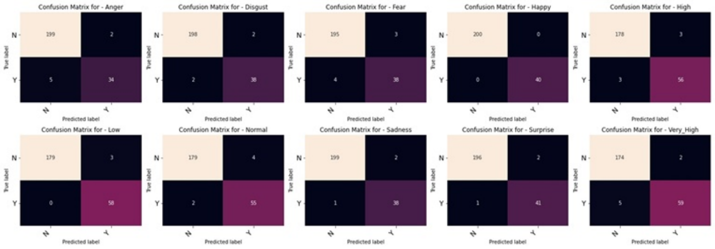


Figure 4.5: Multilabel confusion Matrix for VGGML-CNN Predictions on BU-3DFE.

are presented in Table 4.7. Table 4.8 is the comparison results of the proposed model with some existing models.

Manifold GCN performance on the datasets in this work is presented in Table 4.7. The evaluation is based on the accuracy and loss, ROC AUC, F1_Score and average precision of the test samples. The training and the validation accuracy and loss are shown in Figure 4.7. The comparison study with some existing models is presented in Table 4.8, and the model was compared with some baseline methods on the datasets using accuracy as the metrics for performance evaluation. The pro-

Table 4.3: The result of the comparative studies of multilabel models' performances on BU-3DFE

ML-Models	Hamming Loss ↓	Ranking Loss ↓	Average Precision ↑	Coverage↓
RAKELD [41]	0.4126	0.6859	0.2274	4.8137
CC [42]	0.1807	0.8393	0.3107	4.8094
MLkNN [43]	0.1931	0.8917	0.2634	4.9486
MLARAM [44]	0.3045	0.6552	0.3180	3.1970
ML-CNN	0.1273	0.2867	0.5803	2.5620
VGGML-CNN	**0.0890**	**0.1647**	**0.7093**	**1.9091**

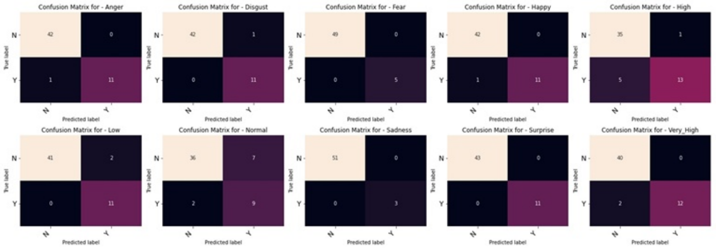


Figure 4.6: Multilabel confusion Matrix for VGGML-CNN Predictions on CK+.

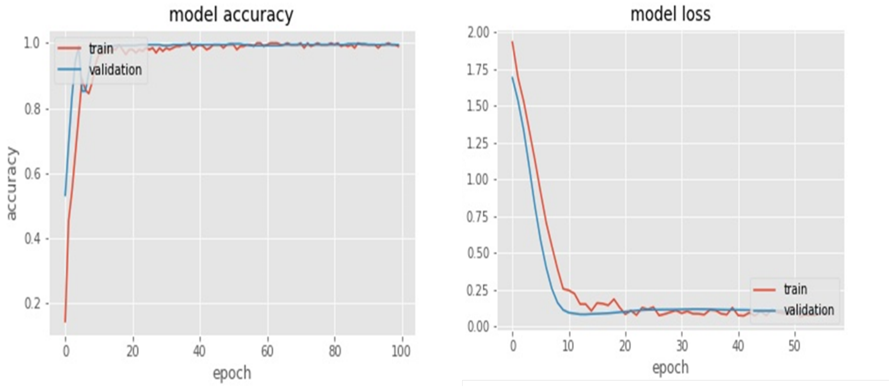


Figure 4.7: Training and Validation Accuracy of Manifold GCN.

posed method has an outstanding performance on BU-3DFE data, given accuracy of 98.80%. The model performance on CK+(97.20) is promising compared to other models. The results implied that similarity measures accounted for by our model, positively affect the system's performance. Figure 4.8 and Figure 4.9 are the confusion matrix predictions of Manifold GCN on the test set of CK+ and BU-3DFE data respectively.

In the experiments in Section 3, investigation for the appropriate value of K (number of neighbour nodes) appropriate for model stability and efficiency was carried out. Increasing the number of k parameters was considered to observe it's

Table 4.4: The comparative studies of multilabel models' performances on augmented BU-3DFE dataset

ML-Model	Hamming Loss ↓	Ranking Loss ↓	Average Precision ↑	Coverage↓
RAKELD [41]	0.3858	0.7223	0.2241	4.0453
CC [42]	0.1825	0.8948	0.2812	4.7270
MLkNN [43]	0.1929	0.9025	0.2573	4.9623
MLARAM [44]	0.3169	0.6963	0.3280	2.9315
ML-CNN	0.1124	0.2278	0.7216	2.2397
VGGML-CNN	**0.0628**	**0.1561**	**0.8637**	**1.3140**

effect on the model performance. K values are increased in the multiple of two, that is, set {2,4,6,8} as presented in Table 4.9. It was observed that the model gives stable and optimal results for the datasets when k= 4. One could deduce that the more the neighbour size, the more complex the classification task. The complexity is suspected to be a result of accommodating members with different features or members belonging to another class, introducing information that misguides the central node classification.

Graph Convolutional network was implemented as a semi-supervised model, couple with manifold enhancement, the model can learn correlation among labels and resolves data inconsistencies. The model's output is the softmax activation function. Softmax approximately gives the proportion of each emotion in the expression face, and removing the threshold enables the model to recover the emotion distribution from logical labels. Figure 4.10 shows the Model's softmax layer predictions on a CK+ test sample. The information learned about nodes and neighbours in the model was visualised and this is presented as points on the graph. Since the model's output holds 16 vectors, then 16-dimensional space is required for the pictorial view. However, we employ Isomap manifold multidimensional reduction to view the data distributions in a manifold plane. The colours represent the classes of the expression data. Figure 4.11 shows the embedding of BU-3DFE and CK+ respectively. Our model learns the correlation among labels from the figure, as the figure reveals that different colours surround some nodes.

4.5 Conclusion

This Chapter has critically and logically presented the results of the frameworks discussed in Chapter 3. Deep forest performance evaluation on CK+ gives a performance accuracy above 90% but less than 70% accuracy on BU-3DFE. Also, the confusion matrices presented in Figure 4.2 indicate the efficiency of deep forest in recognising emotions. CK+ confusion matrix showed that deep forest performed least in recognising sad emotion; it misclassified angry for sadness. The BU-3DFE confusion matrix showed that deep forest performance was best in recognising happy and surprise emotion, average in recognising disgust, fear and sadness and not good

Table 4.5: Comparison of results of VGGML-CNN with some recent models on CK+

Author	Model	Accuracy %	No of Classes	Target
Cai et al. [46]	IL-CNN	94.35	7	Expression only
Li and Deng [27]	DLP-CNN	95.78	7	Expression only
Mohammad et al. [47]	DBN-GSA	98	7	Expression only
Xu et al. [45]	CCNN	91.50	6	Expression & intensity
Chen et al [34]	LDL-ALSG	93.08	7	Expression distribution
Proposed	VGGML-CNN	**97.16**	6	Expression & Intensity

in recognising angry as it misclassified angry for disgust. Figure 4.3 and Figure 4.4 showed the ability of a multilabel convolutional network model to recognise emotion with the associate ordinal intensity concurrently. VGGML-CNN results presented from Table 4.2 through to Table 4.6 showed optimal results and better performance than any of the multilabel algorithms considered. Visualising the output of manifold GCN on CK+ and BU-3DFE showed that the model learnt the correlations among data annotations. According to the results presented in this Chapter, Manifold GCN performance was outstanding on BU-3DFE and promising on CK+ based on accuracy. The summary of the models' performance based on accuracy is presented in Table 4.10

Table 4.6: Comparison result of VGGML-CNN with some recent models on BU-3DFE

Author	Model	Accuracy %	No of Classes	Target
Fernandez et al. [36]	FERAtt	85.15	7	Expression only
Shao and Qian. [48]	ResNet	88.70	6	Expression only
Bao et al. [49]	CNM	80.63	6	Expression only
Proposed	VGGML-CNN	**98.01**	6	Expression and intensity

Table 4.7: The Performance of the Manifold GCN model on test set of the datasets.

Model	Database	Accuracy	F1 Score	Average Precision	Average Recall
M-GCN	BU-3DFE	98.80%	0.97	0.97	0.98
	CK+	97.20%	0.96	0.96	0.97

Table 4.8: The Performance of the manifold GCN model compare with existing models on CK+ data.

Author	Method and contribution	Number of classes	Accuracy
Chen et al [34]	LDL-ALSG	7	93.08
Liu et al.[50]	CDLLNET	7	87.42
Wu et al[51]	E3DNET	7	80.36
Gan et al.[52]	MA-FER	7	83.28
Li and Deng[53]	DBMNET	7	85.27
Ekundayo et al. [54]	Deep Forest	6	93.07
Proposed Model	**Manifold GCN**	**6**	**97.20**

Table 4.9: Effect of kneighbour size on Manifold GCN .

DATABASE	K = 2	K = 4	K = 6	K = 8
BU-3DFE	94.03%	99.30%	87.53%	80.16%
CK+	94.72%	95.05%	86%	72.59%

Table 4.10: Comparing the performance of the proposed models on CK+ and BU-3DFE.

DATABASE	Deep Forest	VGGML-CNN	Manifold GCN
BU-3DFE	65.53%	98.01%	98.80%
CK+	93.22%	97.17%	97.20%

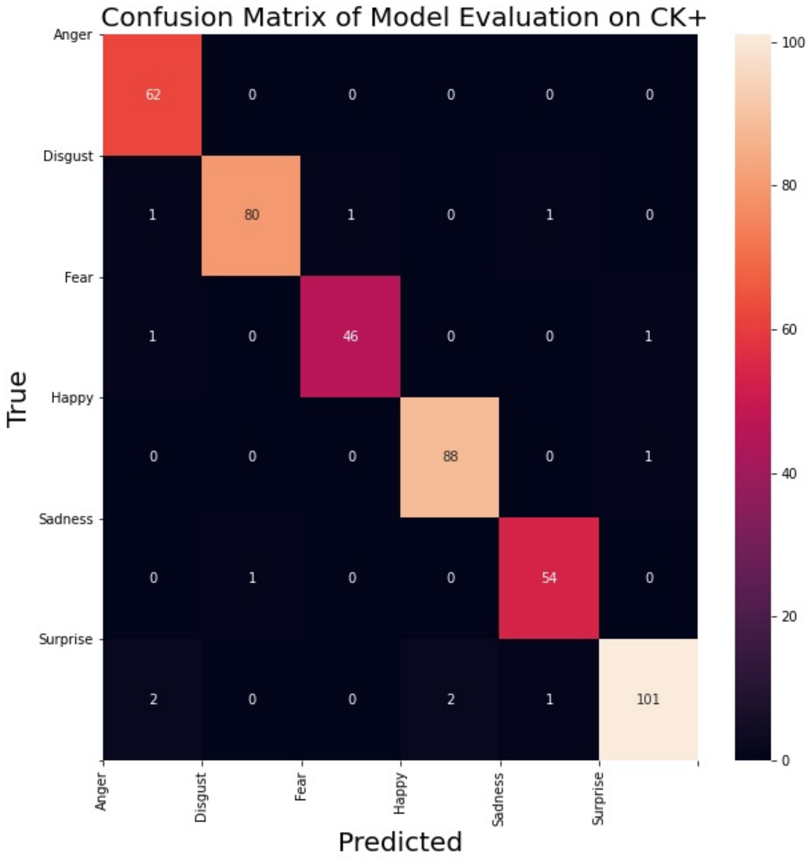
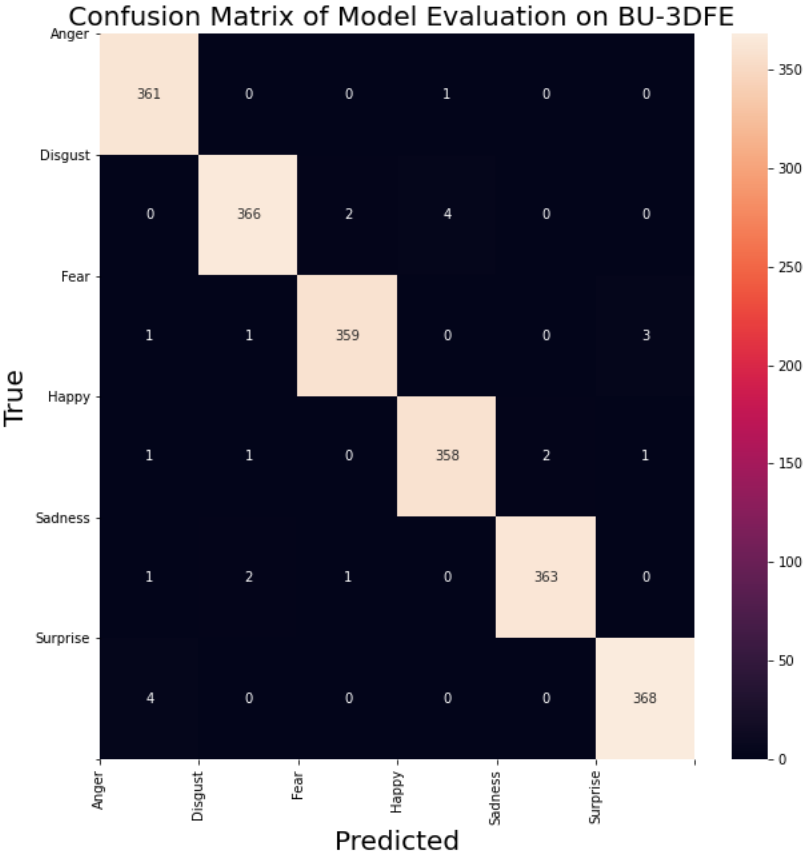


Figure 4.8: Confusion Matrix of Manifold GCN prediction on test set of CK+ data.



	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	361	0	0	1	0	0
Disgust	0	366	2	4	0	0
Fear	1	1	359	0	0	3
Happy	1	1	0	358	2	1
Sadness	1	2	1	0	363	0
Surprise	4	0	0	0	0	368

Figure 4.9: Confusion Matrix of Manifold GCN prediction on test set of BU-3DFE data.

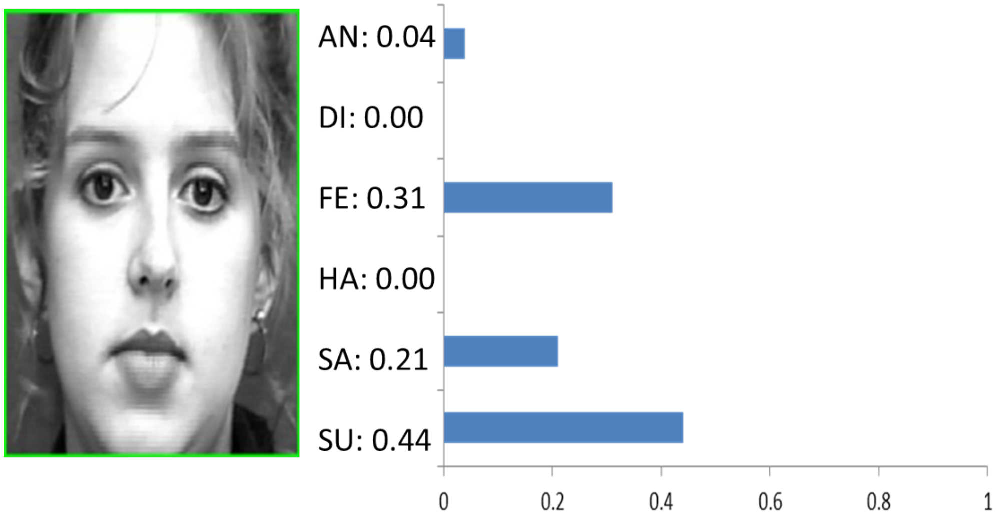


Figure 4.10: The distribution prediction of Manifold GCN softmax output layer on a CK+ sample test [2]

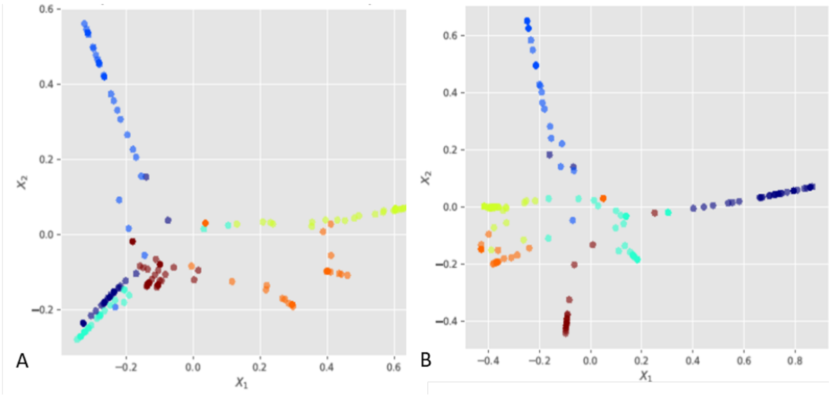


Figure 4.11: BU-3DFE (A) and CK+ (B) Manifold learning visualisation

Chapter 5

Conclusion and Future work

This thesis focused on the facial expression recognition tasks by providing some frameworks for the understanding of emotion recognition techniques improvement. Frameworks for expression recognition, concurrent emotion recognition with intensity estimation and label enhancement techniques for efficient facial expression recognition were presented. The results obtained from the proposed methods have been critically and thoroughly discussed. The rest of this chapter will include an overview of the discussions in the previous chapters, their contributions to knowledge, and suggestions of possible future works.

5.1 Conclusion

This thesis commenced with a comprehensive introduction of facial expression recognition and intensity estimation concepts and provided detailed information on the common challenges and the study's objectives.

The first objective investigated the current trends and techniques studies on FER, especially in emotion recognition and intensity estimation, data annotation inconsistencies and correlation among labels. The outcome presented diverse facial expression recognition applications and discussion of single-label learning, multilabel learning and label distribution learning trends. Single label learning described FER as a multiclass problem, while multilabel learning resolved data annotation ambiguity and label distribution learning, considered data annotation inconsistencies by studying the correlation among labels. Each of these models was thoroughly elucidated together with their limitations. The Chapter also reviewed the techniques popularly used in the field for emotion recognition, ranging from preprocessing techniques, handcrafted techniques, machine learning techniques and state-of-the-art deep learning techniques. These methods were critically and systematically analysed, showing their strengths and limitations.

The second objective investigated deep forest performance in a facial expression recognition environment. An investigative study that aimed at achieving layer-by-layer learning with a shallow learner algorithm; the shallow learner algorithm was a forest tree. The result captured using confusion matrix and accuracy, revealed the deep forest framework's performance on CK+ and BU-3DFE datasets and found that deep forest provided competitive results with CK+ but the performance degraded with more challenging data (BU-3DFE). Deep forest achieved 93.22% accuracy on CK+ and 65.53% on BU-3DFE.

Another objective investigated the multilabel definition of FER in terms of emotion and intensity recognition using ordinal metrics. The multilabel framework employed binary relevance for data transformation and convolution neural network as a classifier with a sigmoid output layer. The multilabel framework was targeted towards recognising emotion from facial images with the respective intensity. Observation showed that the model could successfully recognise emotion with its degree of intensity. Comparing the multilabel convolution neural network with standard multilabel models showed an optimal performance on CK+ and BU-3DFE data. Despite the performance of the multilabel convolution neural network, the model was susceptible to overfitting. Multilabel convolution neural network performance was improved by minimizing model overfitting and intraclass variation and maximizing interclass variation using a pretrained VGG-16 network and island loss.

VGGML-CNN accuracy showed outstanding performance with an accuracy of 97.16% and 98.01% on CK+ and BU-3DFE respectively. VGGML-CNN also showed outstanding results with multilabel metrics (hamming loss, ranking loss, average precision and coverage error) than RAKELD, CC, MLkNN, MLARAM. Multilabel confusion matrices that showed the recognition rates of each emotion and ordinal intensities were also presented.

The last objective was to study the correlation among data labels and resolve data annotation problems using manifold learning and graph convolutional networks. This framework presented the facial expression task as a graph related problem. The manifold algorithm modelled the correlation among data annotations in terms of similarity distance. Each data point is considered a node of the graph and the similarity distances generated by the Isomap manifold are considered the graph's edge information. GCN as a semisupervised learner, used the information of neighbouring nodes to predict the central node. Manifold GCN could recover the label distributions of the emotion from logical labels, resolve label ambiguity and label inconsistency. The Manifold GCN results showed that the two datasets' training and validation accuracy and loss graphs revealed that the model correctly learnt the appropriate feature for emotion classification without overfitting. Manifold GCN output was visualised with Isomap dimensionality reduction. The graph showed that Manifold GCN learnt the correlation among labels because of the clustering of nodes of different colours.

5.2 Contribution to Knowledge

The following are the specific contributions of this study to knowledge.

- A comprehensive review that captured different areas of facial expression recognition applications and also presented trends of facial expression recognition research, categorising them into Single Label Learning (SLL) / Multiclass, Multilabel Learning (MLL) and Label Distribution Learning (LDL).
- A multilabel convolution neural network framework for concurrent emotion recognition and intensity estimation with ordinal metrics.
- Improved the performance of the Multilabel CNN model with the pre-trained network to prevent overfitting, and the island loss function for intraclass variation minimisation and interclass variation maximization.
- Resolved data annotation inconsistency and recovered label distribution from logical label using Manifold Graph Convolutional Network.

5.3 Future work

Facial expression recognition frameworks presented in this work proved their worth. Nevertheless, there are some aspects where they could be improved. This section highlighted some areas of future work to improve the models' performance.

Deep forest performance investigation on facial expression data showed that the performance degraded as data size increased and when challenged data was employed. This limitation could be attributed to fine-grained features used in the model. The limitation could be addressed by employing deep features at the front phase of the model. The combination of deep features and deep forest for a robust facial expression recognition system should be investigated in future work.

This study is conducted in a controlled environment, that is, on laboratory prepared data. Future work should consider more challenging environments like data in the wild or dynamic environments.

The computation of distance similarities in this study (manifold learning) is an inductive process in isolation from the semi-supervised learning GCN, such that GCN could only make predictions based on the prior knowledge of seen data. This limitation could be improved by integrating the algorithms to exhibit inductive learning.

List of References

[1] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3D facial expression database for facial behavior research. *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216. x, 8, 124, 127

[2] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 94–101, San Francisco, CA, USA, 2010. IEEE. x, 8, 128, 135

[3] Ashraf Abbas A. Al-modwahi, Onkemetse Sebetela, Lefoko Nehemiah Batleng, Behrang Parhizkar, and Arash Habibi Lashkari. Facial expression recognition intelligent security system for real time surveillance. In *World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP'12)*, pages 1–8. WorldComp, 2012. 1

[4] Ayesha Butalia, Maya Ingle, and Parag Kulkarni. Facial expression recognition for security. *International Journal of Modern Engineering Research (IJMER) www.ijmer.com*, 2(4):1449–1453, 2012. 1

[5] Hudnall Stamm. Clinical applications of telehealth in mental health care. *Professional Psychology: Research and Practice*, 29(6):536–542, 1998. 1

[6] Diederich Joachim and Insu Song. Mental health informatics: current approaches. *Studies in Computational Intelligence*, 491:247–253, 2014. 1

[7] Swarup Poria, Ananya Mondal, and Pritha Mukhopadhyay. Evaluation of the intricacies of emotional facial expression of psychiatric patients using computational models. pages 199–226. Springer Science + Business Media, 2015. 1

[8] Jens Uwe Garbas, Tobias Ruf, Matthias Unfried, and Anja Dieckmann. Towards robust real-time valence recognition from facial expressions for market research applications. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 570–575. IEEE, 2013. 1

[9] Gozde Yolcu, Ismail Oztel, Serap Kazan, Cemil Oz, and Filiz Bunyak. Deep learning-based face analysis system for monitoring customer interest. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):237–248, 2020. 1

[10] Christine Lisetti and Diane Schiano. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & CognitionPragmatics and Cognition*, 8(1):185–235, 2000. 1

[11] Paul Ekman and Wallace V Friesen. Constant across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. 1

[12] Brais Martinez and Michel Valstar. *Advances , Challenges , and opportunities in automatic facial expression recognition.* Springer International Publishing Switzerland 2016, Switzerland, 2016. 1, 2

[13] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaBoost and support vector machines. *Sensor*, 13:7714–7734, 2013. 1

[14] Na Du, Feng Zhou, Elizabeth Pulver, Dawn Tilbury, Lionel Robert, Anuj Pradhan, and Jessie Yang. Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation Research Part C: Emerging Technologies*, 112:78–87, 2020. 1, 4

[15] Bhavin Sheth and Thuan Pham. How emotional arousal and valence influence access to awareness. *Vision Research*, 48(23):2415–2424, 2008. 1, 4

[16] Yong Yang and Yue Sun. Facial expression recognition based on arousal-valence emotion model and deep learning method. In *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, pages 59–62. IEEE Xplore, 2017. 1, 4

[17] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 1

[18] Sourav Dey Roy, Mrinal Kanti Bhowmik, Priya Saha, and Anjan Kumar Ghosh. An approach for automatic pain detection through facial expression. *Procedia Computer Science*, 84:99–106, 2016. Proceeding of the Seventh International Conference on Intelligent Human Computer Interaction (IHCI 2015). 2

[19] Zhanli Chen, Rashid Ansari, and Diana Wilkie. Automated detection of pain from facial expressions: A rule-based approach using aam. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 8314:125–, 02 2012. 2

[20] Weitong Guo, Hongwu Yang, Zhenyu Liu, Yaping Xu, and Bin Hu. Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Frontiers Neurosci.*, 15:342, 2021. 2

[21] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 248–255. IEEE Xplore, 2017. 2, 4

[22] Robert Plutchik and Jesse M Bering. Integration, differentiation, and derivatives of emotion. *Evolution and congnition*, 7(2), 2001. 2, 4

[23] Yanpeng Liu, Yuwen Cao, Yibin Li, Ming Liu, Rui Song, Yafang Wang, Zhigang Xu, and Xin Ma. Facial expression recognition with PCA and LBP features extracting from active facial patches. In *2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016*, pages 368–373, Angkor Wat, Cambodia, 2016. IEEE. 3

[24] Cigdem Turan and Kin Man Lam. Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study. *Journal of Visual Communication and Image Representation*, 55(January):331–341, 2018. 3

[25] Olufisayo Ekundayo and Serestina Viriri. Facial expression recognition: A review of methods, performances and limitations. In *2019 Conference on Information Communications Technology and Society, ICTAS 2019*, pages 1–6. IEEE, 2019. 3

[26] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. 3, 4

[27] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 4, 130

[28] Russell James and Lanius Ulrich. Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology*, 4(2):119–135, 1984. 4

[29] Siti Khairuni, Amalina Kamarol, Mohamed Hisham, Heikki Kälviäinen, Jussi Parkkinen, and Rajendran Parthiban. Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. *Pattern Recognition Letters*, 92:25–32, 2017. 4

[30] Ragini Verma, Christos Davatzikos, James Loughead, Tim Indersmitten, Ranliang Hu, Christian Kohler, Raquel E. Gur, and Ruben C. Gur. Quantification of facial expressions using high-dimensional shape transformations. *Journal of Neuroscience Methods*, 141(1):61–73, 2005. 4

[31] Michel Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transaction on Systems, Man, And Cybernetics-Part B: Cybernetics*, 42(1):28–43, 2012. 4

[32] Hiroki Nomiya, Shota Sakaue, and Teruhisa Hochin. Recognition and intensity estimation of facial expression using ensemble classifiers. *International Journal of Networked and Distributed Computing*, 4(4):203–211, 2016. 4

[33] Changqin Quan, Yao Qian, and Fuji Ren. Dynamic facial expression recognition based on K-order emotional intensity model. In *2014 IEEE International Conference on Robotics and Biomimetics, IEEE ROBIO 2014*, pages 1164–1168. IEEE, 2014. 4

[34] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13981–13990. IEEE, 2020. 4, 127, 130, 132

[35] Oscar Luaces, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4):303–313, 2012. 61

[36] Pedro Marrero Fernandez, Fidel Alejandro Guerrero Peña, Tsang Ing Ren, and Alexandre Cunha. Feratt: Facial expression recognition with attention net. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, 06 2019. 124, 131

[37] Walid Hariri, Hedi Tabia, Nadir Farah, Abdallah Benouareth, and David Declercq. 3d facial expression recognition using kernel methods on riemannian manifold. *Engineering Applications of Artificial Intelligence*, 64:25–32, 2017. 124

[38] Dmytro Derkach and Federico M. Sukno. Local shape spectrum analysis for 3d facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 41–47. IEEE, 2017. 124

[39] Anwar Saeed, Ayoub Al-Hamadi, Robert Niese, and Moftah Elzobi. Frame-based facial expression recognition using geometrical features. *Advances in Human-Computer Interaction*, 2014:1–13, 04 2014. 124

[40] Md. Zia Uddin, J. J. Lee, and T.-S. Kim. An enhanced independent component-based human facial expression recognition from video. *IEEE Transactions on Consumer Electronics*, 55(4):2216–2224, 2009. 124

[41] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2011. 125, 128, 129

[42] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine*

Learning and Knowledge Discovery in Databases, pages 254–269. Springer, 2009. 125, 128, 129

[43] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007. 125, 128, 129

[44] Benites Fernado and Sapozhnikova Elena. Haram: A hierarchical aram neural network for large-scale text classification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 847–854. IEEE, Nov 2015. 125, 128, 129

[45] Ruyi Xu, Jingying Chen, Jiaxu Han, Lei Tan, and Luhui Xu. Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. *Computing*, 102(3):765–780, 2020. 127, 130

[46] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James Oreilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 302–309, Xi'an, China, 2018. IEEE. 130

[47] Wael Mohammad, Alenazy Abdullah, and Saleh Alqahtani. Gravitational search algorithm based optimized deep learning model with diverse set of features for facial expression recognition. *Journal of Ambient Intelligence and Humanized Computing*, 2020. 130

[48] Jie Shao and Yongsheng Qian. Multi-view facial expression recognition with multi-view facial expression light weight network. *Pattern Recognit. Image Anal.*, 30:805–814, 2020. 131

[49] Wenzheng Bao, Yifeng Zhao, and Deyun Chen. A facial expression recognition method using capsule network model. *Scientific Programming.*, 2020:805–814, 2020. 131

[50] Tingting Liu, Jixin Wang, Bing Yang, and Xuan Wang. Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. *Infrared Physics and Technology*, 112:103594, 01 2021. 132

[51] Zhan Wu, Tong Chen, Ying Chen, Zhihao Zhang, and Guangyuan Liu. Nir-expnet: Three-stream 3d convolutional neural network for near infrared facial expression recognition. *Applied Sciences*, 7(11), 2017. 132

[52] Yanling Gan, Jingying Chen, Zongkai Yang, and Luhui Xu. Multiple attention network for facial expression recognition. *IEEE Access*, 8:7383–7393, 2020. 132

[53] Shan Li and Weihong Deng. Blended emotion in-the-wild : multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127:884–906, 2018. 132

[54] Olufisayo Ekundayo and Serestina Viriri. Deep forest approach for facial expression recognition. In *PSIVT 2019 International Workshops*, pages 149–160, Sydney, NSW, Australia,, 2019. Springer. 132