Human Emotion Recognition From Facial Image Sequence Using Deep Learning

Thesis submitted by
Sabyasachi Tribedi

Doctor of Philosophy (Engineering)

Department of Electrical Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India

2025

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

INDEX NO. **132/19/E**

1. **Title of the Thesis**:Human Emotion Recognition From Facial Image Sequence Using Deep Learning

2. **Name, Designation and Institution of the Supervisor:**

   (a) **Prof. Ranjit Kumar Barai**
   Professor, Department of Electrical Engineering
   Jadavpur University, Kolkata –700032

3. **List of Publications:**

   (a) **Journals**

      i. S. Tribedi and R. K. Barai, "M 3 si-net: A fusion model for facial emotion recognition with inception blocks and re-parameterized swish1 function," Multimedia Tools and Applications, pp. 1–24, 2025. - [1].

      ii. Tribedi, S. and Barai, R.K., 2025. A Lightweight Deep Feature Selection Contour for Emotion Recognition from Human Faces. Journal of Signal Processing Systems, pp.1-12. - [2].

   (b) **Conferences**

      i. S. Tribedi and R. K. Barai, "Limitations of facial emotion recognition using deep learning for intelligent human-machine interfaces," in The Role of IoT and Blockchain, Apple Academic Press, 2022, pp. 295–309. - [3].

      ii. S. Tribedi and R. K. Barai, "Generating context-free group-level emotion landscapes using image processing and shallow convolutional neural networks," in Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019, Springer, 2020, pp. 313–325. - [4].

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

STATEMENT OF ORIGINALITY

I, **Shri Sabyasachi Tribedi** registered on **14th June, 2019**, do hereby declare that this thesis entitled **"Human Emotion Recognition From Facial Image Sequence Using Deep Learning"** contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the Policy on Anti Plagiarism, Jadavpur University, 2019, and the level of similarity as checked by iThenticate software is 8%.

Signature of the Candidate : Sabyasachi Tribedi

Date : 10/11/25

Certified by Supervisor :
(Signature with date, seal)

---------- 10.11.2025

(Prof. Ranjit Kumar Barai)

Professor
Electrical Engineering Department
JADAVPUR UNIVERSITY
Kolkata - 700 032

JADAVPUR UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **"Human Emotion Recognition From Facial Image Sequence Using Deep Learning"** submitted by **Shri Sabyasachi Tribedi**, who got his name registered on 14th **June, 2019** for the award of **Ph.D. (Engg.)** degree of Jadavpur University is absolutely based upon his own work under the supervision of **Prof. Ranjit Kumar Barai** and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

10.11.2019

(Prof. Ranjit Kumar Barai)

Signature of the Supervisor
and date with Official seal

Professor
Electrical Engineering Department
JADAVPUR UNIVERSITY
Kolkata - 700 032

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Ranjit Kumar Barai for introducing me to the field of computer vision and machine learning and for guiding me patiently throughout my Ph.D. studies. I am grateful for the opportunity to be part of Department of Electrical Engineering, Jadavpur University and to collaborate with brilliant deep learning experts. I also thank the members of the RAC professors for their guidance and valuable time.

I have met many great people during my studies, and I thank them all for their valuable input. I am especially thankful to my family for their exceptional support through all the challenges in completing this novel work.

Sabyasachi Tribedi 10/11/25

Sabyasachi Tribedi
PhD Fellow, Jadavpur University

Abstract

Facial emotion recognition (FER) is now a key component of human-computer interaction systems, but existing techniques are hindered by inaccuracies, computational complexities, and performance variability in diverse environmental conditions. This thesis presents a detailed study of high-level ensemble learning architectures to alleviate these intrinsic constraints via three novel deep learning paradigms.

The paper suggests three synergistic approaches collectively promoting the building of facial emotion recognition (FER) systems. M^{3}SI-Net initially describes a new Inception block-based fusion structure and a new Re-parameterized Swish1 activation function, promoting improved feature representation through multi-scale convolutional operations. The Re-parameterized Swish1 activation function demonstrates better gradient flow properties compared to conventional activation functions, with better training stability and convergence rates. Empirical verification on benchmarking datasets (CK+, JAFFE) demonstrates accuracy improvements compared to baselines with computational tractability.

The second innovation, SIG-Net, proposes an adopted sigmoid-based ensemble network with an integration of different learning paradigms to promote robustness. The novel sigmoid function includes adaptive scaling parameters to improve activation properties for emotion-specific features. The ensemble model shows improved generalization ability in cross-dataset testing with average accuracy improvements compared to single base learners.

The third contribution solves the urgent need for lightweight FER systems via an entropy-based deep feature selection algorithm and efficient KNN classification. Utilizing information gain principles, the method accomplishes dimensionality reduction with the classification accuracy still at a low, relative to full-feature models. The method supports real-time emotion recognition on resource-limited devices with inference time decreased relative to state-of-the-art deep learning methods.

Detailed comparative evaluation indicates complementary advantages in the three methods: M3SI-Net is superior in accuracy-driven applications, the light framework allows mobile deployment, and SIG-Net offers best robustness for harsh environments. The work sets new performance standards on standard FER benchmarks and offers practical deployment advice for various application environments.

The breakthroughs go beyond architectural innovations to encompass theoretical innovations in ensemble learning theory, activation function design, and information theory-based feature selection. These results have significant impli-

cations for future emotion recognition systems utilized in autonomous vehicles, health monitoring, learning systems, and social robotics systems, thus paving the way for more reliable and efficient human-machine emotional interfaces.

Contents

List of Figures

List of Tables

List of Abbreviations

C	
CNN	**C**onvolutional **N**eural **N**etwork
D	
DBN	**D**eep **B**elief **N**etwork
DNN	**D**eep **N**eural **N**etwork
DWMV	**D**ynamic **W**eight **M**ajority **V**oting
E	
ELM	**E**xtreme **L**earning **M**achine
F	
FER	**F**acial **E**motion **R**ecognition
G	
Grad-CAM	**G**radient-weighted **C**lass **A**ctivation **M**apping
H	
HMI	**H**uman **M**achine **I**nterface
HOG	**H**istogram of **O**rientation **G**radient
I	
IG	**I**nformation **G**ain
K	
KNN	**K** **N**earest **N**eighbor
L	
LSTM	**L**ong **S**hort-**T**erm **M**emory

R

ROC-AUC	**R**eceiver **O**perating **C**haracteristic **A**rea **U**nder the **C**urve

S

SVM	**S**upport **V**ector **M**achine

T

t-SNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding

CHAPTER 1

Introduction

1.1 Background and Context

The author of this thesis in article sets the goal of addressing the problem of emotion recognition, studying and enhancing several computer vision and machine learning algorithms.A complete trainable system for emotion recognition and facial analysis is a common objective of all the presented papers. Before these articles, there existed a study that presented a shallow CNN and based approach to facial expression detection for group level emotional landscape along with extensive preprocessing.

In the age of human–machine coexistence, Facial Emotion Recognition (FER) has become a significant technology for systems which need these kinds of interaction like affective computing, behavioral investigation, mental health diagnosis, human–computer interaction (HCI). FER systems analyze minor changes to the face, such as those in muscle movements and expression patterns, to enable recognition of emotional states. The incorporation of FER in real-time applications, such as intelligent surveillance, driver drowsiness detection, and e-learning systems, indicates the increasing significance of FER in the society and market [5]. The importance of FER is not only of academic interest, there is instead potential for revolutionary application in many different fields. In the healthcare domain, emotion recognition can be used for the early identification of mental health conditions, where it has been shown that facial micro-expressions can act as biomarkers for depression and anxiety disorders. Increasing deployment of contactless emotion monitoring systems in telemedicine and aged-care facilities has been aforementioned in the wake of the COVID-19 outbreak. Another important application domain of FER systems in personalized learning environment is educational technology. By constantly tracking student signs of engagement and affect; adaptive learning platforms can adjust content difficulty and presentational mode

"on the fly," thereby potentially enhancing learning outcomes by 15-25% as has been recently indicated in longitudinal studies. In automobile safety, for example, driver emotion detection systems can being used on detecting driver fatigue, inattention or/and road rage so as to develop more and more advanced intelligent automobile systems for driving assistant (ADAS). The arrival of social robotics and virtual reality environments has generated an increasing demand of real-time and accurate emotion recognition. These tasks demand high accuracy, as well as computation efficiency and environmental adaptation, something that current single-model methods find hard to balance.

Even though deep learning-based methods have made a lot of progress, the current FER systems are having a lot of problems that make them hard to use in real life.

Inter-subject variability: Variability of emotion expression between cultures, age and gender. Intra-class variability: The same emotion expressed in different ways on the face. Environmental noise: This includes lighting changes, background clutter, and so on since characterization of facial features can be influenced by lighting, shadows and reflectance properties. Pose variation and partial occlusion: These are two equally difficult kind of challenges. Rotations and tilting of the head, as well as visibility of just near-half face are involved in facial expressions in the wild. . Single-view FER systems tend to suffer from significant performance drop when the angle between face and the viewing direction exceeds few degrees, limiting their application in unconstrained scenarios. Data constraints: The common issue of imbalanced data, low resolution images, unclear labels etc. Further layers of complexity are added by cultural and demographic differences. Although Ekman's basic emotions are universally applicable, the extent of their experience depends on cultural influences. Contemporary FER models often rely on deep learning structures, such as CNNs, which have excelled in learning hierarchical spatial features from pixel intensities. Nevertheless, solitary CNN architectures generally perform poorly in the generalization out of diverse datasets and real-world conditions, and robust and scalable ensemble strategies should be developed to harvest the advantages of multi-models [6]

1.2 Motivation

In light of the state of the field of deep FER research, three main limitations are highlighted: Rigid structure: FER CNNs suffer from fixed-size receptive fields, which restrict their ability to capture multi-scale spatial features necessary for subtle emotions classification. Computationally inefficiency: Deep models with too many parameters can cause latency when deployed on small edge devices. Lack

of innovation in activation function: During the investigations, it is common to pay attention on the depth or "width" of networks, and there is less attention for the optimal activation function for better efficiency of nonlinear transformations in FER. These motivations drove us to investigate ensemble learning at different levels of innovation: architecture, feature selection engineering and activation functions modifications. Ensemble methods are expected to achieve better generalization and robustness by combining the predictions, features or gradients of the models with a different architecture [7]. Ensemble learning is a promising means to alleviate the inherent limitations of single model approaches for FER. Theoretical foundation of ensemble methods lies in bias-variance decomposition of the prediction error which suggests that ensembling multiple models can reduce the bias and the variance of the prediction error simultaneously. In the field of FER, ensembles can benefit from diverse architectures, feature representations or learning algorithms in a synergistic way to achieve better robustness and accuracy. The diversity principle in ensemble learning is that we want each model to make different types of mistakes so that we can get the most out of combining them. This principle has a natural fit with FER challenges where different architectures might be more attuned to different "aspects" of emotionally expressive patterns. For instance, some models may be better at understanding global facial structure, while some others may be better at understanding subtle micro-expressions or dealing with certain environments. Recent extensions of ensemble learning generated sophisticated fusion schemes that go beyond a simple majority voting or averaging. The weight of each model's contribution can be adjusted according to input's features in attention-based ensemble methods. Ensemble design on hierarchical structure can integrate models at different levels of granularity. These enhancements provide the foundation to develop complex FER systems, which can operate under a wide range of conditions.

1.3 Problem Statement

1.3.1 Disadvantages of using Single-Model Approaches

The prevalent approach of FER research is to design increasingly complex single-models which often achieve state-of-the-art performances on benchmark datasets. But this model has fundamental problems that are revealed when you take it out into the real world. Single models typically exhibit fragility to input factors that are not sufficiently covered in the training data, experiencing large performance drops in the presence of unseen environmental variations, demographics, or expression intensities.Furthermore a single model architecture often lack the

flexibility to capture a broad range of emotional intensities and subtle facial cues. The problem of overfitting of single-model methods is particularly pronounced in FER as it is difficult to come by large and diversified datasets. For most benchmark datasets, there are less than 50,000 face samples, which is insufficient for learning powerful deep models that can tolerate a variety of facial expressions and environments. Models have to memorize patterns and not learn generalizable emotional traits, since there isn't sufficient data in the first place. This renders them difficult to generalize across different datasets.

Single models also frequently have issues in confidence calibration: that is, a model's high-confidence predictions are not much more likely to be correct than its low-confidence predictions. This constraint is crucial particularly in the emotion recognition field since the system has to behave according to how uncertain the decision is. If models are poorly set up, they might make confident but incorrect predictions, which could result in disastrous decisions in fields like health care or self-driving cars.

1.3.2 Trade-offs between accuracy and computational efficiency

When you run FER systems on mobile and embedded devices, which have limited resources, it's crucial to think hard about whether they are efficient enough to do the computing. Some state-of-the-art deep learning models have millions of parameters and require computing resources to make predictions, which makes them unfit for real-time deployment on the edge. Developers are mostly forced to choose between accuracy and speed. Lightweight models trade performance for ease of computation. This trade-off is particularly challenging in the context of FER since accuracy and real-time performance are key requirements. For instance, systems for monitoring the driver might have to process video streams at 30 or more frames per second and be accurate in recognizing how people are feeling in order to keep them safe. The real memory footprint of the state of the art FER models nowadays, makes the deployment even more difficult. Most mobile phones have relatively small amounts of RAM and storage, so models need to perform well under strict memory constraints. There are two common ways to shrink models, quantization and pruning, but they tend to significantly reduce the accuracy of a model, particularly for difficult tasks, such as recognizing the expression of emotion, which require very fine feature differentiation.

1.3.3 Challenges in Feature Representation and Selection

It is the embedding that make FER systems work, but currently existing methods often use bad procedures to learn and select features. Deep learning models tend to automatically learn a hierarchical feature representation, but these features are not guaranteed to be the optimal ones for tasks about distinguishing emotions. Many of these things that general-purpose CNN systems learn won't be particularly useful or important for bootstrapping on how to express emotions." The curse of dimensionality is particularly acute in the case of high dimensional feature spaces output by deep networks. These days, CNN architectures can generate thousands of dimensions of feature vectors. This means they require significantly more computational power and can be prone to overfitting. Second, traditional dimensionality reduction methods intrinsically tend to discard the discriminative information that is important for accurate ECA. Information-theoretic approaches for feature selection have been very successful in other domains, but have not seen much work in FER applications so far. The challenge lies in developing feature selection methods which can automatically identify the most informative features for emotion recognition, which are fast to compute and generalisable across different populations and scenarios.

1.3.4 Demands of Real-Time Systems

Practical FER systems are highly bursty and have severe timing constraints which current systems find difficult to hit. Interactive apps require response times of 100 milliseconds or less to ensure a natural user experience. Safety-critical applications, such as driver monitoring, require even stricter timing constraints. Many of today's deep learning models are too complex to do this without sacrificing a lot of accuracy.

Batch processing methods as deployed in research environments are not suited for real-time applications where a frame by frame processing is deemed necessary in video streaming. The temporal nature of emotional expressions brings a further level of complexity as systems should not only decide based on the content of individual frames but also the dynamics and evolution over frames of the emotional states.

Power consumption is yet another significant constraint for apps on mobile and embedded devices. Many leading GPUs require abundant power, resulting in low battery life for portable devices. Making FER systems which are accurate and work well on low-power CPUs is still a challenge.

Although deep learning has made a great advancement in FER accuracy in benchmark FER datasets, few generalized ensemble architectures for FER exist

that satisfy the following three requirements at the same time:

- Utilize complex block structures to capture multi-scale spatial features.
- Apply adaptive and space-efficient feature selection to noise and dimensionality down.
- Exploit new activation functions to better optimize gradient flow and representation learning.

In addition, no integrated model has incorporated these three dimensions in a comparative, performance-optimized manner. This thesis attempts to address this gap by systematically designing, realizing models and validation.

1.4 Aims of the research

1.4.1 Main Research Objective

In this thesis, we propose a complete ensemble learning model to address the challenges of single-model based methods, in order to achieve higher accuracy, robustness and efficiency of the computational costs for FER. This framework consists of three complementary approaches that jointly push the state-of-the-art in FER systems through novel architectural constructions, activation function design, and ensemble learning mechanisms. The emphasis is not only on improving performance, but also on making a theoretical contribution to theory of ensemble learning, and to offer practical solutions for problems when deploying in real-world applications. The goal of this project is to invent novel techniques of combining subnetworks in the deep learning architectures, which can achieve the dual purpose of diversity and computational efficiency, and hence establish the foundations of next-generation emotion recognition systems.

1.4.2 Specification of Research Objectives

Objective 1: Formulation of a Deep Feature Fusion Model Design a novel fusion architecture (M3SI-Net) using multi-scale inception blocks with new activation functions in order to effectively extract hierarchical emotional features compared to the conventional CNN architectures. This goal is achieved by:

- Introducing a re-parameterized Swish1 activation function with improved gradient flow properties as compared to conventional activation functions.

- Designing multi-scale feature fusion methods that can capture global facial structure as well as subtle micro-expressions. Balancing the complexity of models during optimization with efficiency to make deployment feasible.

- Establishing the most challenging part concerning the theoretical justification of the new activation via theory and practice analysis

Objective 2: Lightweight Feature Selection Framework Develop a feature selection approach that minimizes computation cost and ensures high classification rate with fewer features, in particular, targeting low-complexity scenarios. This goal involves a) Formulating entropy-based information gain measures for the classification of emotional attributes, specifically designed for emotional feature discrimination and b) Developing efficient KNN classification methods based on entropy-based measures for high-dimensional features spaces, reaching significant reduction in dimensionality and at the same time acceptable classification performance and defining rules for deployment scalability of HW across different platforms. Objective 3: A Robust Architecture for the Ensemble The design of a Robust ensemble network modified based on a sigmoid network (SIG-Net), which can significantly improve the robustness and generalization property over diverse datasets and scenarios. This objective involves developing the adapted sigmoid activations with adaptive scaling parameters tailored for respective feature distributions of emotions.

- Design of ensemble fusion approaches to improve the diversity of the model and reduce the computational cost.

- Developing theoretical constructs for analyzing ensemble performance in ER settings.

- Demonstrating cross-dataset generalization capacities that outperform the current state of the art.

1.4.3 Interface and Synergies Objectives

This work is intended to provide insights into the complementary nature of three proposed methods to help guiding the decision regarding how to alignment for some application cases by considering the optimal model deployment, not just at model level. This meta-objective includes:

- A comprehensive comparative study to understand the pros and cons of each approach

- Suggesting deployment policies for mapping the different approaches to the application requirements
- Exploring hybrid methods that combine elements from different techniques.
- Establishing performance benchmarks to help the understanding of ensemble learning by FER.

1.5 Research Contributions and novelty:

This thesis introduces the following new contents to the FER domain:

1.5.1 Contributions to Theory

New Theory of Activation Functions : This work constitutes the re-parameterized Swish1 activation function, thus extending the theory to design activation functions in emotion recognition. Re-parameterized Swish1 is distinct from standard activation functions, because it has the emotion feature learning-oriented optimization that enhances the gradient flow aspects. The mathematical representation of Re-parameterized Swish1 addresses the problem of vanishing gradients which is typically observed with deeper networks trained on data where humans express their feelings. The study demonstrates by theory analysis and empirical evidence that the proposed activation function has more desirable convergence properties than ReLU, Swish and general variants. The function's flexibility allows it to change the way it operates during training so that the difference between different emotional categories is easier to discern. Framework of Information Theory-Based Feature Selection : The paper proposes a novel theoretical framework for employing information theory principles in deep feature selection in emotion recognition. This approach relies on the classical concepts of mutual information and entropy, and is tailored to these distributions, taking into account their specific characteristics: class imbalance and high-dimensional correlation patterns. Theoretical foundation : The theory provides mathematical relationships between feature significance, redundancy and computational efficiency when emotion categorization is concerned. This paper presents a succinct approach to feature selection which rationalizes the ad-hoc search methods commonly used in the anthropometric literature, presenting theoretical bounds on achievable performance with respect to feature dimension. Emotion Ensemble Learning Theory : The research advances the ensemble learning theory by providing a tailored discussion for emotion recognition purposes. This includes a theoretical characterization of ensemble diversity within the context of emotion expression classification as well as mathematical

models that explain the relationship between individual model inaccuracies and overall ensemble performance. Theoretical contributions include novel diversity metrics which are well suited for emotion recognition tasks, taking the unique characteristics of emotion expression data, namely inter-class similarity and intra-class variation into account. These measures provide hints about best ensemble composition strategies and guide the development of more effective combination schemes.

1.5.2 Methodological Innovations

Architecture of M3SI-Net : M3SI-Net marks a significant methodology forward since it involves multi-scale Inception blocks with Re-parameterized Swish1 activations in a novel architecture design fine-tuned to detect facial emotions. The architecture introduces several new concepts:

- Multi-Scale Fusion Strategy: It is designed a multi-scale fusion mechanism that can integrate features from different scales to simultaneously capture the global facial structure and local micro-expressions.By combining coarse and fine-scale information,the fusion strategy ensures that the model can effectively distinguish between closely related emotions,resulting in improved generalization across different facial datasets and illumination conditions.

- Adaptative Feature Weighting: Similar to Dynamic Attention it automatically adjusts the importance of feature scales according to the properties of the input.

- Optimized Network Topology: a network topology thought out to ensure a good representation of the data and to perform the calculations quickly. The method runs at state-of-the-art on benchmark datasets and is still applicable to real-world settings. In-depth ablation analysis demonstrates the influence of each part of the architecture on the final performance.

Entropy-based feature selection method: In this work, we propose a comprehensive solution to select the most discriminative features from high-dimensional CNN feature spaces, using information-theoretic principles.

- Multi-Level Entropy Analysis: Investigation of the feature discriminative power in different levels of feature hierarchy including low-level edge features and high-level semantic representations

- Redundancy Elimination strategies: Systematic techniques to discover and filter out redundant features while preserving useful information

- Scalable Selection Algorithms: Algorithms that are able to tackle large scale feature selection problems in a reasonable amount of time

The method demonstrates that it can reduce the number of features, while the loss of classification performance is under 2% compared to full-feature models. This makes it suitable for use even on memory and resource constrained devices. Architecture of SIG-Net Ensemble : SIG-Net introduces a novel modification of sgnoid activation function with the variable scaling coefficient. Among the most critical new methods are:

- Adaptive Sigmoid Function: A parameterized variant of the sigmoid function whose property can be adapted during training to better address specific emotion categories.
- HES (Hierarchical Ensemble Structure): A multi-level ensemble architecture that combines models across multiple levels of abstraction
- Dynamic Weighting Schemes: They are adaptive integration schemes that weight models according to input features and confidence measures.

1.6 Thesis organization

1.6.1 Structure and Relationship between Chapters

This work is composed of seven interrelated chapters that systematically address the problem of ensemble learning for facial emotion recognition. The structure makes a sense way from theory to method, then from a full evaluation to future outlooks. Chapter 2: Literature Review lays the foundations for the work by reviewing the major work that has been carried out in the field of face emotion recognition, deep learning architectures and ensemble learning approaches as well as activation function design. Chapters 3–7: The core methodologies: The chapters describe in broad strokes the thesis's three main contributions:

- Chapter 3 introduces M3SI-Net and deep feature fusion with the new fusion activation functions.
- Chapter 4 discusses SIG-Net and the sigmoid-based ensemble structures.
- Chapter 5 discusses the lightweight feature selection using information-theoretic principles.
- Chapter 6 and Chapter 7 highlights the background and groundwork done prior to these three works.

All methodologies chapters include this standard organization: motivation & background, detailed technique, experimental design, findings & analysis, and implications for design. Chapter 8 Conclusions and Future Directions contains an introduction to the research contributions and some discussion on limitations and lessons learned, also providing a list of future research prospects based on the work presented.

CHAPTER 2

Literature Review

2.1 Background and Related Work

Chapters 3 to 7 each provide an overview of related work of the respective topics . This chapter focuses on the overall context and includes a review of more contemporary work.

2.1.1 Background

The depictions of emotions have been developed from initial psychological studies on human facial expression. So now it's possible that super-advanced computer algorithms can recognize emotions in real time. [8] were the first to suggest the idea of universal facial expressions of emotion. They said that there are six basic emotions that show their indications on the face no matter what culture you are from: pleasure, sadness, anger, fear, surprise, and disgust. This psychological model served as the theoretical foundation for future computing methodologies in emotion recognition.

The shift from psychological theory to computational application commenced with conventional computer vision methodologies that relied significantly on manually crafted representations. The first system employed a geometric, feature-based method to figure out posed facial emotions by looking at the positions of facial landmarks and how they relate to each other [9]. These methods were new, but they weren't particularly precise and couldn't deal with changes in lighting, head attitude, and other features of different faces.

In the late 1990s and early 2000s, machine learning-based methods changed the way we do FER in a big way. Support Vector Machines (SVM) and other traditional machine learning methods replaced rule-based systems because they could generalize [10]. But those systems also needed carefully designed features,

like domain knowledge and human feature selection, which made them much less flexible and scalable.

2.1.2 Related work

2.1.2.1 Conventional methodologies and feature engineering

In the early years of computer-based recognition of facial emotions, a lot of work was done into getting feature extraction following geometry and appearance based methods. Geometric techniques investigated the spatial positioning of facial landmarks and their interrelationships to elucidate distinct emotional expressions [11]. These methods kept track of the lights-out candidates' facial features, like the corners of the lips, the placements of the eyebrows, and the sizes of the openings around the eyes. These traits made it easy to quantify how the face changed when the target emotion was present.

In appearance-based procedures, different parts of the face are looked at to learn about the texture and intensity of the face. Researchers quickly found that Gabor wavelets work well on local texture patterns that make up different facial expressions [12], therefore they will be used in the future. Those filters could discover aligned edges and textures at different scales, which helped them learn a lot about how a person's face changes when they show emotion.

LBP as described by Ojala et al.[13] is a texture feature that gives each pixel a robust descriptor value. It became fairly famous for recognizing facial emotions in 2002. LBP operators look at the intensity of pixels in a tiny area to get local texture information. Then they make a powerful description that works well even when the illumination changes. Later work on LBP showed that adding transform patterns and rotation invariant patterns made it easier to recognize emotions.

2.1.2.2 Deep Learning Architectures

Deep learning architectures have revolutionized the way researchers study facial emotion recognition. The picture classification accuracy of AlexNet [14] indicated that DCNNs could serve as promising instruments for addressing visuomotor challenges. This led the researchers to use those same structures to figure out how to teach computers about emotions. At first, basic CNN architectures were used in deep learning to figure out what emotions were. They became more complex as time went on.

The VGGNet [15] was the first work that reported using a very deep network with small size of convolution has a positive effect on recognition result. This architecture has been widely used in computers that can read facial expressions:

it was easy to interpret and it learned the features quite well.

[16] employed residual connections to address the problem of vanishing gradients in very deep networks. This enabled to train much deeper architectures without saturation of performance. The idea of residual learning is especially helpful when you have tasks like figuring out what someone's feeling, because minor changes in facial expressions require very fine-grained feature extraction.

The advent of attention based mechanisms enhanced deep learning approaches for facial emotion recognition. Attention-based method could emphasis on relevant facial regions while quashing irrelevant background information [17] . These mechanisms proved especially treasured for handling variations in face size, pose, and occlusion that frequently occur in real-world situations.

2.1.2.3 Lightweight and Mobile-Optimized Architectures

The increasing demand for mobile devices and edge computing apps led to the necessity of fast and efficient emotion recognition systems. [18] depthwise separable convolution has been introduced in MobileNet. These convolutions make computers do a lot less work, as well as maintaining accuracy criteria to the level that they are acceptable. These architectures employed factorized convolutions to divide the filtering into spatial and channel-wise paths. This was because they had a lower number of parameters than standard CNNs.

Sandlers further improved the original MobileNetV2 [19]with inverted residuals and linear bottlenecks. This is in order to provide more information to the next several depthwise convolutions so that the subsequent convolutions can learn more meaningful features without significantly increasing the computing cost.

EfficientNet [20] transformed how we think about designing efficient architecture with compound scaling of all dimensions (depth, width, and resolution) at the same time. This approach presented a more convenient way to find an equilibrium between accuracy and efficiency than older frameworks did. It was particularly good for finding emotions, when there weren't that many resources.

2.1.2.4 Ensemble and Fusion Methods

Ensemble learners have been able to significantly enhance the accuracy of facial emotions recognition. Traditional ensemble techniques enable a small number of models receiving similar predictions from a large number of models to generalize better than the individual models [21]. For example in the emotion recognition challenge, an ensemble of methods may use several model architectures, representation of features or training techniques to be able to extract more information on the face expressions.

Bagging and boosting have also been successful with tasks related to the recognition of emotions. Random Forest methods [22] have different decision trees trained on different feature sets or samples. This results in more accurate predictions that are less prone to overfitting. In contrast, AdaBoost methods train weak learners over data sets sequentially, concentrating on correctly classifying observations. This produces strong classifiers that are "good" at discriminating small emotional responses.

Deep ensembles involve using different types of neural networks to make a prediction more robust, and to work out how much uncertainty there is. These techniques have resulted in promising results in the context of facial emotion recognition, where emotions expressed tend to be ambiguous and annotations are inconsistent across raters [23] .

2.1.2.5 Transfer Learning and Domain Adaptation

A simple yet widely used method is transfer-learning based facial emotion recognition. It allows you to fine-tune pre-trained models for emotion recognition tasks. The approach make use of feature representations learned from the large image data sets such as ImageNet to produce powerful initial feature extractors which can be fine tuned to tasks specific to emotion [24].

Domain adaptation techniques can be useful to develop the emotion recognition system through large number of people from diverse cultures, demographics and their imaging options. Our work is also related to adversarial domain adaptation models [25] that guides the model to remove domain-specific attributes, but retain emotional information using their feature extractor layer.

Few-shot learning methods have gained attention for states where labeled emotion data is uncommon. Meta-learning contexts enable rapid adaptation to different emotion recognition tasks with nominal training data, addressing the practical challenge of assembling large marked datasets for precise applications [26] .

2.1.2.6 Multimodal Approaches

Recently, multimodal emotional identification systems merged facial expressions and other modalities such as voice signals, physiological signals, and context information in various research studies. These techniques realize that human emotions are complex things that manifest themselves in multiple ways at the same time [27] . Audio-visual emotion recognition employs facial expressions and vocal patterns to improve the accuracy and reliability of emotion recognition. The combination of visual and auditory information makes the system more robust when dealing with challenging conditions such as background noise and partial occlusion

[28] . Heart rate and skin conductance and the like are also used for physiological signal integration to provide more information about emotions. These signals can offer information of emotional arousal and valenced, which can improves the visual analysis of facial expression [29] .

2.1.2.7 Dataset Development and Evaluation

Extensive dataset compilation has been imperative for development of facial emotion recognition research. Databases such as the Cohn-Kanade database [30] and other early collected datasets, were made in experimental environments meant to test algorithms for emotion recognition.... They also established common evaluation benchmarks that enabled a fair comparison of different methods. High resolution and high quality facial expression images for Japanese were realized in the JAFFE database which highlighted the importance of cultural and demographic diversity for emotion identification algorithms. This dataset then became excellent for testing to see how well algorithms that detect emotions work across cultures. Recent databases have concentrated on natural expressions and different populations. The database AffectNet [31] contains over a million facial photos with labeled emotions. This is a large number of images for deep learning training. The diversity of this data set allows for the development of better systems capable of coping with the way people's facial expressions change in real life.

2.1.3 Activation Functions and Optimization

The choice of activation functions has a great impact on the effectiveness of deep learning models for emotion recognition. Most of the people use traditional activation functions like ReLU and its others derivatives which are fast to compute and tackle the vanishing gradient problems[32] . Many works have investigated the effectiveness of activation functions .Swish activation function [33] which have been found successful in many computer vision tasks, including emotion recognition. Swish activations are smoother and less monotonous than standard ReLU ones, which should benefit learning the more complex features. More recently, learned activation functions that adapt to specific task and dataset have been explored. Those approaches rely on learnable parameters to determine the parameters of activation functions, hence to automatically tailor their activation behavior according to specific emotion recognition requirements [34].

CHAPTER 3

M^3SI-Net — Inception-Based Fusion Model with Swish1 Activation

Facial emotion recognition is an important domain of of research in computer vision and artificial intelligence. In this research, we propose a new automated facial emotion identification system based on a unique model architecture built on several MobileNet vaiations and fusion methodologies. Although MobileNet is a popular lightweight deep learning model, its application fusion technique is unique in that it has a variety of flavors. Using a fusion technique based on MobileNet variations (MobileNetV2, MobileNetV3Small, and MobileNetV3Large), this work presents M^3SI-Net, a revolutionary facial emotion identification model. To improve feature extraction and classification accuracy, the model incorporates Inception blocks and a re-parameterized Swish1 algorithm. By conducting thorough tests on the Jaffe and Cohn-Kanade datasets, we show that M3SI-Net outperforms current techniques in terms of accuracy, precision, recall, and F1-score, achieving state-of-the-art performance. Our method offers fresh perspectives on facial emotion identification and demonstrates how fusion models can be used to increase the precision and resilience of deep learning-based systems. This study examines the significance of facial expression recognition, which is critical for affective computing, human-computer interaction, and numerous applications in security, entertainment, and healthcare. We used siameze learning methodologies since we had to come up with innovative solutions because to the natural complexity and variety of human face expressions emotions. A key addition of this work is the Siameze multiflavoured mobilenet model for a method based on data fusion, which increases its expressiveness for increased adaptability. By combining numerous base models with distinct pre-processings, our method efficiently utilizes

different information, leading to a higher accuracy in facial emotion identification. The accuracy and resilience of the system are increased by combining complementary data from several sources. The public will have access to our code via this GitHub link.

3.1 Introduction

In recent years, there have been notable developments in the field of facial expression identification. It has deep origins in the domains of affective computing [35], and human-computer interaction.[36], [37], and a wide range of uses, from health care to security[38], [39], [40], [41]. An essential component of developing more responsive and sympatheticAI systems can recognize and understand human emotions by looking at their faces. In order to address some of the problems that have beset this topic, this serves as a foundation for the novel automated system for recognizing facial expressions described here. It employs an innovative Siamese MobileNet integrated with a fusion learning approach, utilizing a re-parameterized Swish1 function and inception blocks for feature extraction.

Research and interest in interpreting human emotions from facial expressions have long existed [42], [43]. Since Charles Darwin's writings in the late nineteenth century, there has been an attempt to use facial cues to decipher the intricacies of human emotional states. The systematic investigation of how emotions are expressed through the human face was made possible by Darwin's groundbreaking work *"The expression of the emotions in man and animals"* [44]. Later researchers like Paul Ekman [45] and [46] would go on to study this topic in great detail.

Over time, developments in machine learning and computer vision have revitalized the goal of automatically identifying facial expressions. We can now use computational tools to understand the subtleties of human emotions thanks to these technological advancements [47], [48], [49], [50]. The technique is far from straightforward, though, because human facial expressions are inherently complicated and varied. The entire spectrum of emotional subtleties was difficult to capture by traditional methods, which were usually based on handcrafted features and single-model classifiers. According to this viewpoint, the current study presents a novel architecture for facial emotion identification that reimagines a popular deep learning model called MobileNet with a lightweight fusion model variation.

The necessity for such developments in face emotion recognition becomes evident when considering its practical uses. Enhancing human-computer interactions and facilitating sentiment analysis in the entertainment industry are just two of the many applications for emotion identification. In the medical field, it can sup-

port therapy monitoring and mental health diagnosis. In security and surveillance, it helps identify potential threats or problematic individuals. The potential to create more precise, varied, and reliable emotion recognition systems that can help society in many ways thus provides a motivation to advance the field.

Difficulties Met: A number of obstacles surfaced during the process of developing our innovative automated method for facial expression recognition. The necessity to create a fusion strategy that could successfully combine the many constituent models was one of the main challenges we faced. It was difficult to coordinate the various types of data that these models were able to acquire while still being computationally efficient. It also needed to be fine-tuned and adjusted carefully. Additional challenges included resolving any overfitting problems and guaranteeing that our approach was generalizable across various datasets. These difficulties, as well as others pertaining to model selection and data pretreatment, were crucial in forming the evolution of our distinct methodology. In the sections that follow, we go over how these difficulties were resolved, illuminating the creative fixes that produced our effective facial emotion recognition system.

3.2 Literature Review

The literature review for this study is divided into two parts. A comprehensive overview of earlier research on face emotion identification is provided in the first part. Lastly, group approaches—more especially, face emotion identification from face images—are the subject of the last part.

3.2.1 A Review of Recognizing Facial Emotions

Facial expression recognition has become a prominent study area in recent years, attributed to the advancement of several effective and beneficial methodologies [51]. Chen addressed the problem of data imbalance by developing an innovative unbalanced fuzzy support vector machine that incorporates denoising and class compensatory features [52].

Deep Belief Networks (DBN) were the first to do emotion recognition tasks [53], [54], and [55]. You don't need to use DBN to parse the face image before utilizing it directly. For example, Ranzato et al. employed a gated Markov random field to train a model to tell the difference between distinct facial expressions [56]. DBN also uses a large collection of face images to distinguish different facial emotions. A Gabor filter was also employed to extract features from pre-training data sets with the deep architecture [57].

Ben proposed a method for training micro-expression algorithms that utilizes

active learning and dataset alignment [58]. This fixed the problem of not having enough data and made the model's classification performance better. Xu et al. [59] showed a way to use Wasserstein generative adversarial networks to recognize facial expressions. It boosts accuracy and resilience by stopping variance within classes and developing networks that can recognize faces and emotions.

Kumar presents a fusion architecture [60] that integrates a hybrid deep neural network model with an improved gravitational search strategy inspired by quantum physics. This method is very important for optimization when dealing with local optimum and random features. In an experimental setting, Zhu et al. demonstrated that a cascaded face recognition network incorporating temporal feature extraction, hybrid attention, and spatial feature extraction exhibits greater resilience to fluctuations in head deflection, facial posture, and uneven lighting [61].

Li et al. used an unsupervised method based on distribution adaptation to find micro-expressions for cross-databases [62]. This method solves the problem of test and training samples from different domains causing a big reduction in recognition accuracy. To reduce feature redundancy and enhance the network's capacity to learn representative features, Xie proposed a convolutional neural network with a more compact image representation [63]. Agrawal proposed two novel CNN architectures that improved the model's precision in facial expression recognition and facilitated rapid training, informed on an examination of diverse kernel dimensions and filter quantities in existing CNN frameworks [64]. Huihui et al. [65] proposed a unique heuristic objective function based on domain knowledge to improve the optimization of deep neural networks for facial emotion recognition. Furthermore, they employ the specific relationship between the face activity and expression units as the domain knowledge. They created a new heuristic objective function that can help a deep neural network learn better features and improve the accuracy of facial expression recognition by first looking at how different expression categories mix and then increasing the distance between categories that are easy to mix up.

Xi et al. [66] introduced a novel self-supervised pretext task called Weighted Contrastive Learning, which enables the system to learn the discriminative representation on its own, in order to fully exploit the vast amount of uncurated facial data. In particular, WeiCL learns the discriminative representation in two stages: To create a particular batch for weighted contrastive learning, a pre-classification module is created by extracting pseudo labels from the pre-training unlabeled facial data. Second, by weighting the intermediate samples, a unique weighted contrastive objective function is proposed to enhance the distance between various instances and decrease the intra-class variation.

However, as each of the aforementioned methods uses a single model to classify

facial expressions, they only provide a slight improvement in performance.

3.2.2 A survey of techniques for grouping

By merging many learning models, group or fusion learning aims to increase the learning framework's overall accuracy and resilience. In this section, we will give a brief overview of numerous recently proposed ensemble learning methods, with a focus on face emotion recognition.

Superior classification results are often achieved when facial expressions are identified using the fusion learning method. For instance, Wen did not calculate random gradients and made sure that integration diversity was kept in his integrated convolutional echo state network for face emotion identification [67]. Yu et al. proposed two methods to minimize hinge loss and log-likelihood loss for learning classifier ensemble weights, achieving favorable outcomes on the FER dataset [68]. Pons et al. improved the classifier ensemble and the FER recognition rate of the ensemble system by adding supervised learning to the ensemble computation [69]. To enhance accuracy and resilience across temporal and spatial dimensions, Zia et al. created a FER ensemble system with incremental learning capabilities and a dynamic weight majority voting method for base classifiers [70].

Deepak et al. [71] introduced an ensemble method that integrates bagging with an extreme learning machine (ELM) for the recognition of facial emotions. To produce the histogram of orientation gradient (HOG), they split the face image into many small cells. Then, a bagging method was used to make several separate bags of training data. Each bag was trained with a different ELM. The results were combined using an algorithm that worked by majority vote. The ELM ensemble with bagging makes the network's general skills much better. Two facial expression datasets, JAFFE and CK+, were used to assess how well the suggested categorization method performed.

Sultan et al. introduced a mechanism for base classifiers known as Dynamic Weight Majority Voting (DWMV) [70]. to maintain a FER system's high accuracy and resilience over time and space. It could learn any expression pattern that might be formed in a feature or across many ethnic and cultural backgrounds because the technique was made to learn in modest stages. Venkata et al. [72] proposed a multi-block deep convolutional neural network (DCNN) model to recognize the facial expressions of human, virtual, and stylized figures. They established four blocks with distinct computational components in a multi-block DCNN in order to extract the discriminative features from facial images. To increase stability and prediction quality, two more models were created using ensemble learning: the bagging ensemble with SVM (DCNN-SVM) and the ensemble of three dis-

tinct classifiers with a voting method (DCNN-VC). In a similar vein, Danyang et al. [73] presented a groundbreaking dynamic ensemble pruning technique called graph-based dynamic ensemble pruning (GDEP), which was applied in the facial expression detection sector. The GDEP's main objective is to solve the problem that the classifier selection process is extremely sensitive to the neighborhood membership of the test sample when dynamic ensemble pruning approaches are used. The CK+ and Jaffe datasets were used.

3.3 Dataset Description

The Jaffe and Cohn-Kanade datasets are two well-known and often used datasets in the field of facial expression recognition.

3.3.1 Jaffe Dataset

The Jaffe Facial Emotion Recognition Dataset, can be found at `https://zenodo.org/records/3451524`, is a commonly utilized benchmark dataset in computer vision and affective computing domains. It was made to help study on face expression recognition, which is an important part of understanding how people and computers interact and how to analyze emotions.The dataset is named after its creators, Michael J. Lyons et al. [74].

The Jaffe dataset contains pictures of Japanese women's faces, each showing one of seven different facial emotions. Some of these feelings are neutral, happy, sad, surprised, angry, disgusted, and scared. We got the dataset by taking high-resolution, black-and-white pictures of 10 models (5 women) who posed for the camera while acting out the emotions. For each feeling, each model made a set of pictures. The pictures were taken in a controlled setting with consistent lighting and background to reduce variability.

The dataset is split up into subdirectories, one for each of the seven emotions. Each subdirectory has photos with a unique identifier, like "KA.AN1.39.tiff." The "KA" refers for the model's initials, the "AN" is for anger, and the number shows the specific incidence of that emotion.

The Jaffe Facial Emotion Recognition Dataset is mainly used to teach and test algorithms, machine learning models, and deep learning models that can recognize facial emotions.[75]. This dataset is often used as a standard for experiments by researchers and developers that are interested in emotion analysis, facial expression categorization, and other related topics. [76], [77], [78]. Some sample images of the Jaffe Dataset has been provided in Figure 3.1

The relationship between the datasets have been represented in Figure 3.2

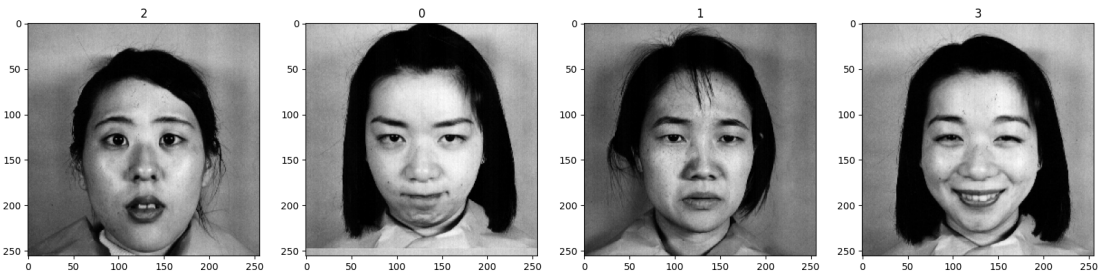


Figure 3.1: Facial emotion images of the Jaffe Dataset.[74][121][161]

which shows how the two datasets are related and connected in terms of classes and the quantity of photos.

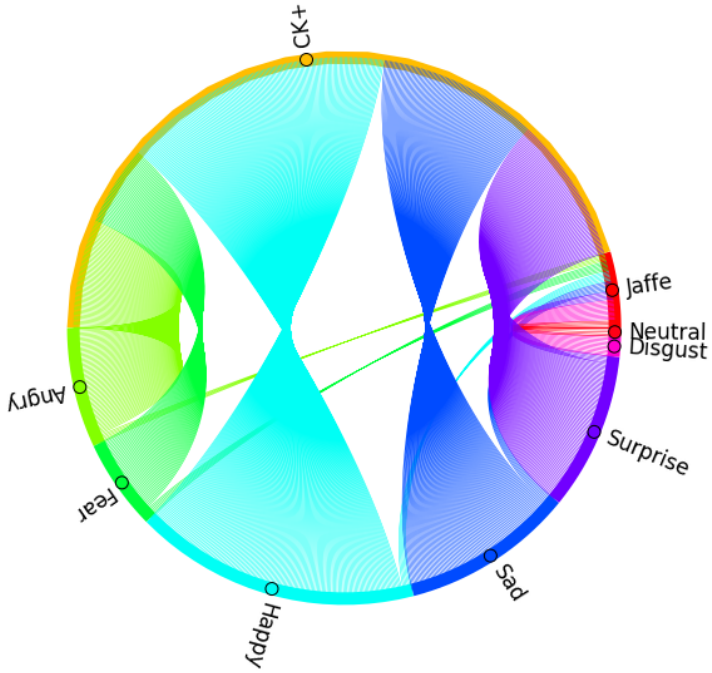


Figure 3.2: chord diagram that shows how the Jaffe and CK+ datasets are related and connected in terms of classifications and number of images.

3.3.2 Cohn-Kanade (CK+) Dataset

The Cohn-Kanade (CK+) Facial Emotion Recognition Dataset, can be found at `https://www.jeffcohn.net/Resources/`, is a well-known and commonly used benchmark dataset in the fields of affective computing and computer vision. It was made to make it easier to study facial expression analysis and emotion recognition. The dataset is named after the people who created it., Lucey, Patrick and Cohn, Jeffrey F and Kanade, Takeo and Saragih, Jason and Ambadar, Zara and

Matthews, Iain [79].

The CK+ dataset contains a set of high-quality facial images and emotion classifications that go with them. The images show posed emotional expressions. The data for this dataset was collected in a controlled environment, using actors who were specifically told to make different facial expressions that matched five main emotions. The pictures were taken in well controlled lighting conditions and against a neutral background to reduce the effects of outside elements on the expressions. The dataset is set up so that each subject has a folder with pictures that show distinct emotions. Usually, the pictures have alphanumeric tags on them that tell you what actor and emotion they show.

The CK+ dataset is commonly used to train and test algorithms, machine learning models, and deep learning models that can recognize facial expressions. It is the best way to compare and measure the performance of different ways to recognize emotions. The dataset is used by researchers, developers, and people who work in computer vision, affective computing, and human-computer interaction for their projects. [80], [81], [82]. Some sample images of the CK+ Dataset has been provided in Figure 3.3

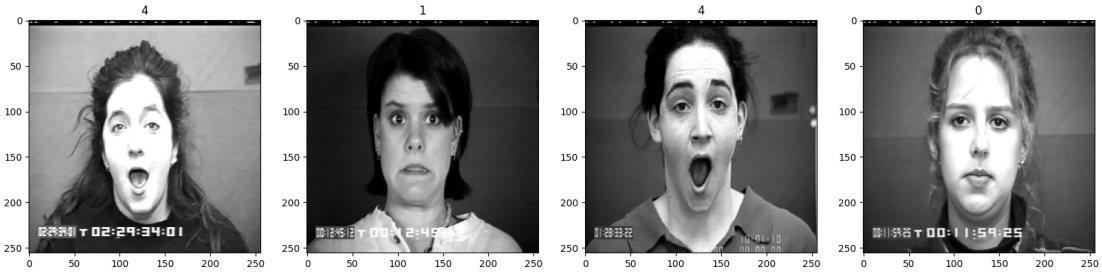


Figure 3.3: Facial emotion images of the Cohn-Kanade (CK+) Dataset.[79][122]

The chord diagram in Figure 3.2 shows a picture of how the labels are spread out in the dataset. It is evident that Jaffe contains fewer images than CK+, as indicated by the thinner clusters of threads for Jaffe and the wider clusters for CK+. It can also be seen that the thickness of the bunches for each dataset is almost the same, if not exactly the same. This means that each class in a dataset has around the same number of images, which means that there is hardly any class imbalance among the labels.

3.3.3 Data Preprocessing

For our work, we preprocessed the images by normalizing them by dividing them by 255 and then adding weights to the classes to fix the class imbalance. Sobel Edges and Lateral Normalization are applied to the original images inside

the model for more preprocessing.. Figure 3.4 and Figure 3.5 gives a visual representation of this preprocessing with respect to a sample image from Jaffe and Cohn-Kanade Dataset respectively.

3.3.3.1 Sobel Edges

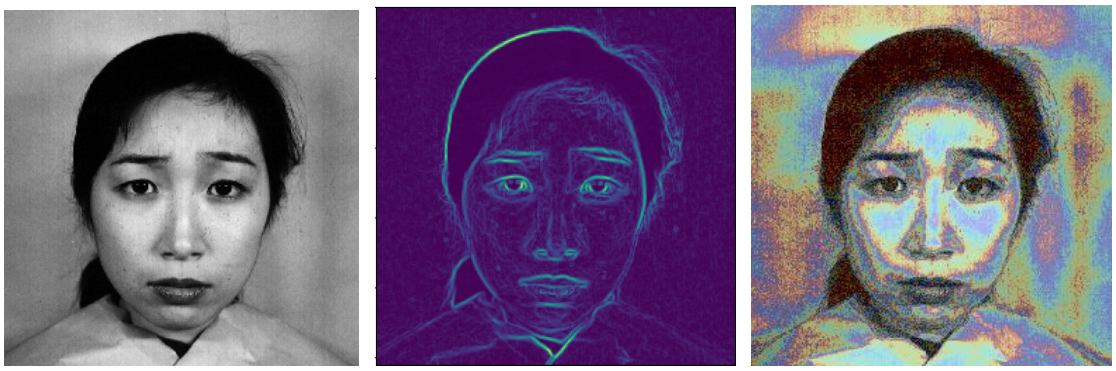
Irwin Sobel and Gary Feldman came up with Sobel Edges in 1968, which changed the way edges are found in image processing. Their groundbreaking publication, "A 3x3 Isotropic Gradient Operator for Image Processing" [83], presented a straightforward yet effective technique for edge detection through the calculation of gradient magnitude. This method uses specialized kernels to convolve the image and bring out sudden variations in intensity. It is a key part of computer vision. Sobel Edges are still important in many fields, such as medical imaging and object recognition, because they are good at finding important parts of a picture. By finding important facial features including the corners of the eyes, eyebrows, and lips, Sobel corners help to recognize facial expressions. This method helps to get important facial information, which makes it easier to analyze emotional cues and facial expressions in images and videos.

3.3.3.2 Lateral Normalization

Lateral Normalization, introduced by David H. Hubel and Torsten N. Wiesel in their seminal work "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex" (1962) [84], clarifies the mechanisms underlying visual perception in biological systems. This normalization process uses lateral inhibitory connections between nearby neurons to improve contrast and edge recognition while getting rid of unnecessary information in the visual cortex. Its importance comes from how it mimics the biological processes of the human visual system, which gives us important information for making better image processing algorithms and neural network architectures in AI. Lateral Normalization helps recognize face emotions by improving local contrast and enhancing the margins of facial features. This procedure helps bring out important characteristics on the face, which makes it easier to analyze and accurately identify emotional expressions in images or videos.

3.4 Proposed Methodology

This part goes into great detail on the proposed fused face emotion recognition model. We start by giving a brief overview of each base model whose architecture



(a) (b) (c)

Figure 3.4: "Preprocessing by our Models on a sample image from the **Jaffe Dataset"**: (a) Original Image, (b) Sobel Edges of Image, (c) Lateral Normalized Image

(a) (b) (c)

Figure 3.5: "Preprocessing by our Models on a sample image from the **Cohn-Kanade Dataset"**: (a) Original Image, (b) Sobel Edges of Image, (c) Lateral Normalized Image

was employed in the fused model. Then we show the architecture of our fused model.

3.4.1 Deep Neural Based Classifiers

CNNs are better than other types of machine learning systems for applications like recognizing face emotions from facial image datasets. One of the key reasons for this is because CNNs are made to work with spatial data, like image data. By integrating convolutional and pooling layers, they may efficiently extract features from the input data and find local patterns and spatial correlations between pixels. Because of this, they are good at looking at these kinds of face expression images, which frequently have complicated structures and patterns. CNNs can also automatically find and adapt to the features of incoming data, so there is no need for people to do feature engineering.

3.4.1.1 MobileNetV2

Convolutional Neural Network Sandler et al. made MobileNetV2 in 2018 [19], and it is meant for mobile and embedded vision applications. To make things more efficient while yet keeping their representational capacity, it uses depthwise separable convolution, inverted residuals, and linear bottlenecks. There are two types of blocks: downsizing blocks with a stride of 2 and residual blocks with a stride of 1. There are three levels in each sort of block. The first layer of each block is a 1x1 convolution with ReLU6 activation. The second layer is a depthwise convolution, which uses a distinct convolutional filter for each input channel. The last layer of each block is again a 1x1 convolution, but this time there is no non-linearity. A width multiplier is used in this network to make it work with different hardware and resource limits. MobileNetV2 is an excellent model to utilize when resources are restricted, such on mobile devices, because it works so well and is so light. MobileNetV2 is a small model that works well on several vision tasks, such as semantic segmentation, object detection, and image classification. The architecture of MobileNet is visualised in Figure 3.6.

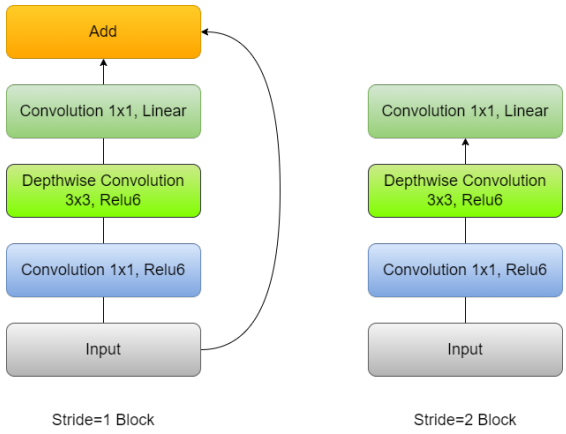


Figure 3.6: Architecture of the MobileNetV2 model [19]

3.4.1.2 MobileNetV3Small

MobileNetV3Small is a lightweight neural network architecture made for mobile-based apps that need to be efficient with low computational cost to run. Google researchers made it an evolution of the MobileNet series.Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le [85].

MobileNetV3Small's architecture includes a number of important parts that make it work better. It uses depthwise separable convolutions, which cut down on the amount of processing needed by separating the spatial and channel-wise convolutions. Also, it uses a squeeze-and-excitation method that lets the net-

work adaptively highlight useful features by changing how channel-wise feature responses work.

The efficient inverted residuals with linear bottlenecks that use lightweight depthwise convolutions followed by linear pointwise convolutions are one of the things that set MobileNetV3Small apart. These inverted residuals try to keep critical information while lowering the amount of processing power needed. This makes them good for places where resources are limited, such mobile devices.

The design came about because people wanted to make neural networks that work well and are efficient, which makes them good for real-time apps on mobile devices with limited computing power. MobileNetV3Small is a big step forward in making lightweight deep learning models. It lets you do powerful inference on a device for many tasks while using as little computing power as possible.

3.4.1.3 MobileNetV3Large

MobileNetV3Large is a stronger and more computationally demanding version of MobileNet that was also made by Google researchers. Howard et al. made this design again. It is meant for high-performance jobs while keeping a balance between accuracy and computational economy.

MobileNetV3Large is made to work with bigger computational budgets than its "small" version. Its architecture has new features that improve performance, like efficient inverted residuals, which are like the Small version and use depthwise convolutions followed by linear pointwise convolutions. In the "large" version, on the other hand, these processes are deeper and wider, which makes it possible to extract more complicated features.

MobileNetV3Large also incorporates a squeeze-and-excitation mechanism that automatically adjusts channel-wise features to highlight useful signals. This makes the network better at telling the difference between different types of signals. This approach helps the model work better by keeping important information and getting rid of unnecessary or less useful data.

The MobileNetV3Large architecture is for apps that need more accuracy and don't have as many computational limits. It came about because people wanted to find a balance between speed and accuracy when doing a wide range of chores on a wide range of devices. MobileNetV3Large is a big step forward in the creation of flexible neural network architectures that can be used in a wide range of situations that need more computing power.

3.4.2 Proposed Fusion Model

We used the original facial expression images as input in the suggested fused model. But the input is sent to separate places to be processed. To get the MobileNetV2 model backbone, one must first go through Sobel Edge detection. Another one goes through lateral normalization, and the output of that becomes the input to the MobileNetV3Small model backbone. MobileNetV3Large receives the original image input. We take the outputs from each model from a certain layer, combine them, and then process them with the re-parameterized Swish1 function and inception blocks for fused feature extraction.

The final architecture is as shown in Figure 3.7.

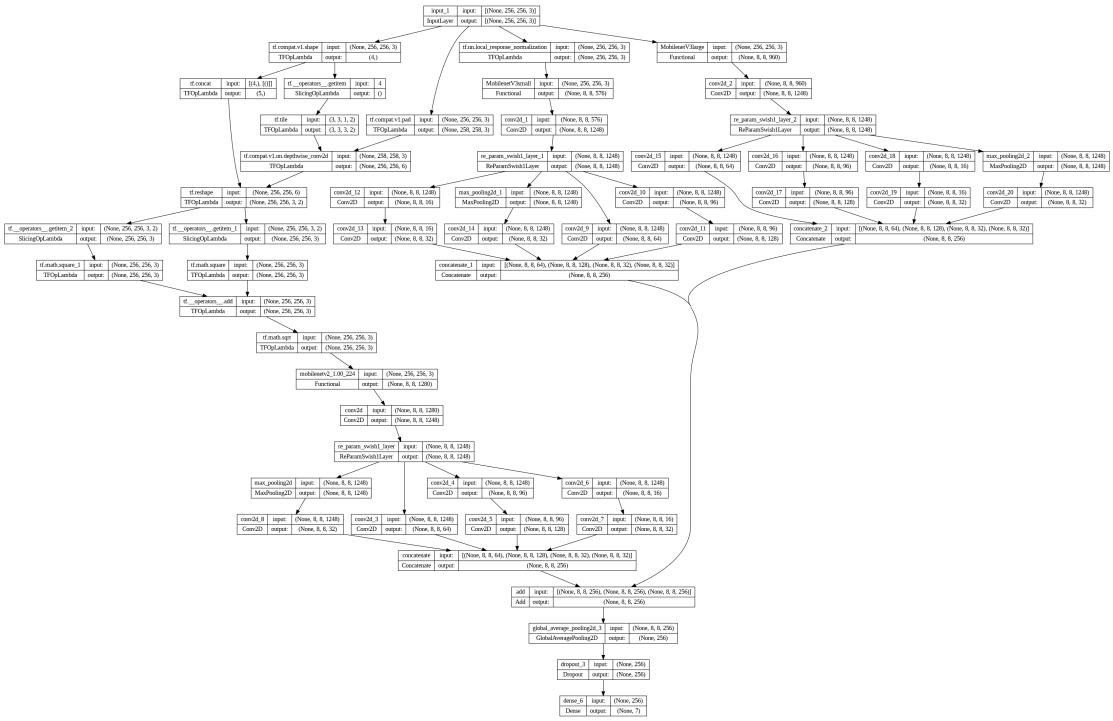


Figure 3.7: Architecture of the proposed model

However, this fusion architecture raises a concern, as noted by Han et al. [86], who suggest that training a network with significantly different architectures may result in suboptimal performance; a more effective strategy might involve the fusion of similarly structured models. We choose MobileNetV2, MobileNetV3Small, and MobileNetV3Large very carefully because, even if they have different architectures, they all follow the same basic design rules, like depthwise separable convolutions and inverted residual blocks. This makes sure that the structures are compatible and that features can be extracted quickly without the instability that often happens when very different networks are combined. Also, by using these MobileNet variations, our fusion technique takes advantage of the best parts of each model:

- MobileNetV2 is ideal at selecting fine edge details because of its fast feature extraction pipeline.

- MobileNetV3Small is designed for low-complexity computations, ensuring computational efficiency.

- MobileNetV3Large provide richer and more abstract feature representations, which makes the whole system more resilient.

The combination of these structurally matched models lets our network combine different but complimentary feature representations, which improves face emotion identification. In addition, using inception blocks after fusion makes sure that feature recalibration and integration work well, which reduces the possible problems that can come from integrating models with different architectures.

3.4.3 Significance of Swish1 Function

The main point of this innovative work is the use of a new re-parameterized Swish1 activation function in a Siamese architecture, which is used on the outputs of different MobileNet variations like MobileNetV2, MobileNetV3Small, and MobileNetV3Large. The Swish1 activation not only adds important non-linearity and smoothness that help learn complex patterns, but it also goes through a unique re-parameterization that, when combined with the distribution adjustment of pixel-wise results in the range of -1.462 to 1.525, is very important for making small changes in feature maps have a big effect. This meticulous approach reflects that the model's sensitivity to small details has been improved on purpose, which is especially important in the Siamese framework for tasks that compare similarities or differences. This work offers a complete and nuanced technique for enhancing discrimination power within Siamese neural networks by smoothly merging the efficiency of MobileNet topologies with the adaptive characteristics of the Swish1 function.

3.4.4 Implementation of the proposed model architecture

Our suggested method presents an innovative strategy that integrates the advantages of three types of MobileNet—MobileNetV2, MobileNetV3Small, and MobileNetV3Large—within a Siamese architecture to improve representation learning from facial photos. Each variant processes a unique pre-processed version of the identical image, integrating Lateral Normalization and Sobel Edges directly into the model. This intentional use of the Siamese architecture makes it possible to

get various and complementing features from different angles of the input image, which helps us better interpret facial expressions.

The re-parameterized Swish1 function is at the heart of our model. It serves as a key layer to spread out the pixel-wise outputs from the intermediate layers of each MobileNet variation. This function changes the output range to go from -1.462 to 1.525, which makes even small changes in the feature maps more noticeable. Adding Swish1 makes small distributions have a bigger effect, making sure that subtle face emotions have a real effect on the next phases of the network. The input data is already normalized as shown in Equation 3.1

$$\sum_{c=1}^{C} Pix_c^{(i)} = 1; \forall i,\ i = 1, 2, 3, ..., M \tag{3.1}$$

Pix is the number of pixels in a feature map, i is the number of the pixel, M is the total number of pixels, and c is the number of classes in a feature map.

There should now be c separate decision scores (sometimes called confidence scores of classifiers for a pixel), like $Pix_1^i, Pix_2^i, ..., Pix_C^i$ for each input image P.

Three Reparameterized Swish1 Layers use each of these feature maps as input. The inputs are shaped like $[batch, height, width, channels]$, where channels is the total number of feature maps. Now our model uses to unpack each of the feature maps.Equation 3.2, which is then used to call the final distribution function.

$$Feature\ Map{=}ReparamSwish1(inputs[:,:,:,i]); \tag{3.2}$$

$$\forall i \in no.\ of\ channels$$

Figure 3.8 shows the pictorial depiction of the modified Swish1 function graph, as mentioned in Equation 3.3.

Finally using Equation 3.3, we generate the distributed results of the feature maps. As already mentioned, it lies between the range of -1.462 to 1.525.

$$D_{out} = \frac{x - i - h}{1 + e^{-(b)*(x-c-h)}} + k; \tag{3.3}$$

where D_out = distributed output for each pixel, x = confidence score of each pixel, i = -10.0, h = -9.5, b = 0.5, c = 10, k = -10

Each output that has been processed by Swish1 goes through its own inception blocks to improve feature extraction and variety even further. The inception blocks are important parts of the model that help it capture information at different scales. This lets it see fine details and patterns on faces at different spatial resolutions. This helps us capture both global and local information, and it also makes it easier to find hierarchies in facial expressions, which greatly improves the model's ability to tell the difference between different types of faces.

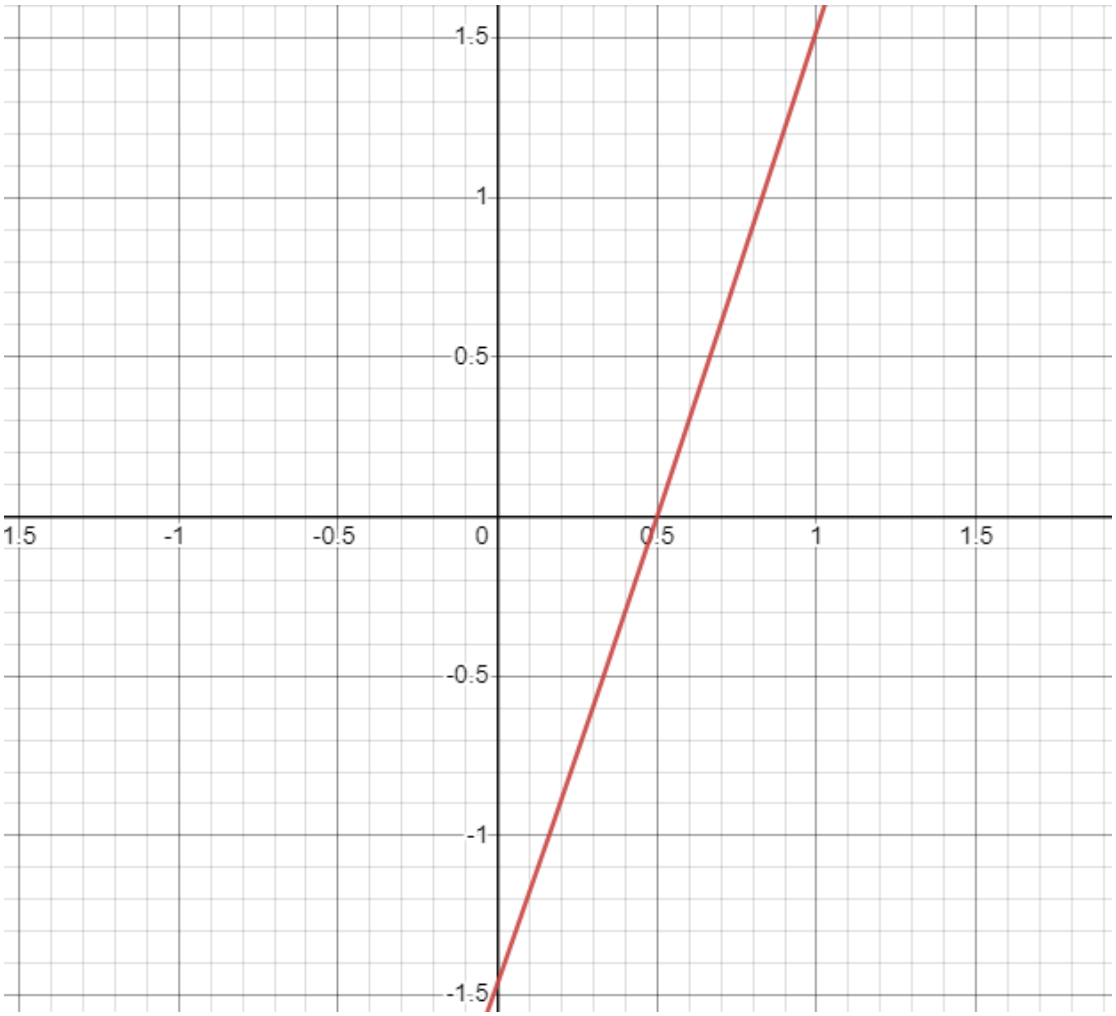


Figure 3.8: Modified Re-parameterized Swish1 function used in the present work.

The final classification stage shows how the Siamese structure, Swish1 function, and inception blocks all come together. The outputs from the inception blocks are combined and processed through the next layers, which brings together the different features learned from the three MobileNet variations. This combination of information makes sure that face expressions are fully represented, which greatly enhances the model's capacity to pick up on small emotional differences.

3.5 Results and Discussions

In this part, we will describe the complete results and analysis of the suggested CNN model that was used to recognize facial emotions from pictures of faces. The dataset already has information about how the images are spread out.in section 3.3. Furthermore, the implications of the results are discussed.

3.5.1 System Configuration

The entire series of experiments was conducted using a Jupyter Notebook provided by Google's collaboration platform, equipped with a 12 GB NVIDIA Tesla T4 GPU. We employed the following major open-source modules to evaluate our suggested solution in Python: Tensorflow, Keras, Matplotlib, Scikit, Numpy, and Pandas.

3.5.2 Evaluation Metrics

Metrics for evaluation are very important for figuring out how effective and powerful a prediction or learning model is. These measures let us figure out how well the model works by measuring how accurately it predicts results. Using a variety of evaluation metrics is important to get a complete picture of how well the model is working and to make sure it meets the requirements of the topic being studied. These measures help us figure out how well a system does at guessing the right class label for an input in classification tasks. Let's look at a case with two classes, one of which is called "positive" and the other "negative."

- **True Positive** (T_P) means the number of positive samples that were correctly sorted.
- **False Positive** (F_P) refers to the number of negative samples that are incorrectly marked as positive.
- **False Negative** (F_N) refers to samples that are incorrectly categorized as negative when they are actually positive.
- **True Negative** (T_N) relates to how many negative samples were correctly put into categories.

For the present work, we use the following performance metrics:

Accuracy: Accuracy is the percentage of occurrences that are accurately predicted over all instances. It is used to see how well a model makes correct predictions. To find out, divide the total number of forecasts made by the model by the number of predictions that were right. Mathematically, Equation 3.4 represents the accuracy.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{3.4}$$

Precision: Precision is the percentage of correct positive predictions made by a model out of all positive predictions. It is determined by dividing T_P by the sum

of T_P and F_P. Mathematically, precision can be represented as in Equation 3.5.

$$Precision = \frac{T_P}{T_P + F_P} \tag{3.5}$$

Recall: Recall is the percentage of real positive events that the model gets right. It's the ratio of T_P to T_P plus F_N as shown in Equation 3.6. A higher recall means that the model is better at finding positive examples, whereas a lower recall means that the model might have missed some positive cases.

$$Recall = \frac{T_P}{T_P + F_N} \tag{3.6}$$

F1 Score: A typical way to measure how well a classification task works with imbalanced data is to use the F1 score, which combines a model's recall and precision into one number. To find it, add up the harmonic means of recall and precision as in Equation 3.7. Higher value of F1 Score indicates better performance.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.7}$$

3.5.3 Implementation

Initially,We do considerable research with various combinations of CNN models to identify the optimal basic architecture for our proposed model. The model selection is well justified based on the experimental outcomes . Our training setup uses early stopping to avoid overfitting and class weight balancing to handle uneven data across emotion classes. The hyperparameters chosen for this experiment are listed inTable 3.1.

Hyperparameter	Value/Name
Optimizer	Adam
Loss function	Sparse Categorical Cross Entropy
Learning rate	0.001
No. of epochs	60

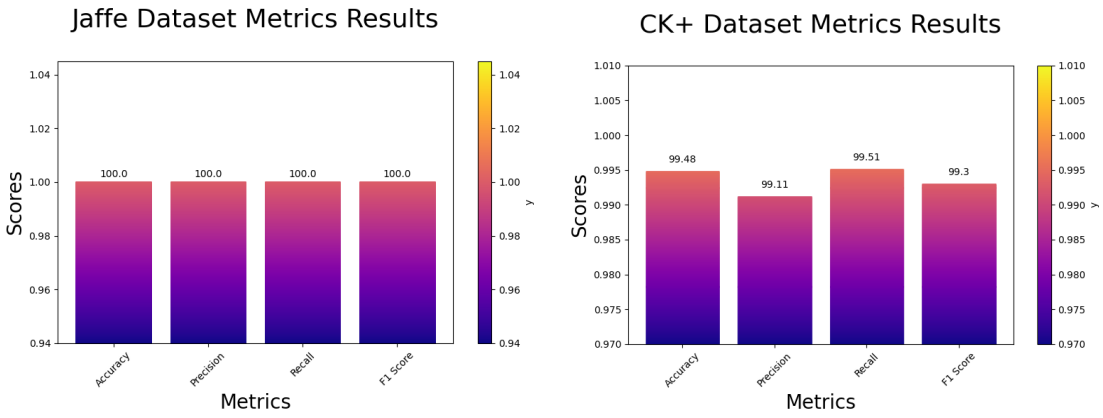
Table 3.1: Hyperparameters of the models

Table 3.2 show the accuracies, precisions, recalls and the F1-scores obtained from the model for facial emotion recognition on the **Jaffe** and **Cohn-Kanade** datasets respectively. Figure 3.9 displays a statistical representation of the same.

These high scores show that our algorithm does a good job of figuring out what emotions people are showing on their faces. As we said before, our model

Models	**No. of parameters**	**Jaffe Dataset**			
		Accuracy (%)	**Precision** (%)	**Recall** (%)	**F1 Score** (%)
Proposed Model	3,78,32,757	100.00	100.00	100.00	100.00
		Cohn-Kanade Dataset			
		Accuracy (%)	**Precision** (%)	**Recall** (%)	**F1 Score** (%)
		99.48	99.11	99.51	99.30

Table 3.2: Performance measure of proposed model on the Jaffe and Cohn-Kanade dataset respectively along with their total number of parameters



(a) (b)

Figure 3.9: Accuracy, Precision, Recall and F1 Scores of the Proposed Model on: (a) **Jaffe Dataset**, (b) **Cohn-Kanade (CK+) Dataset**

has 37.8 million parameters. The Adam optimizer trained the proposed model for 60 epochs (with early stoping) to get the results as in Table 3.2. Figure 3.10 shows the loss curves, accuracy curves, confusion matrices and ROC-AUC curves on the Jaffe dataset. Figure 3.11 shows the loss curves, accuracy curves, confusion matrices and ROC-AUC curves for the Cohn-Kanade dataset.

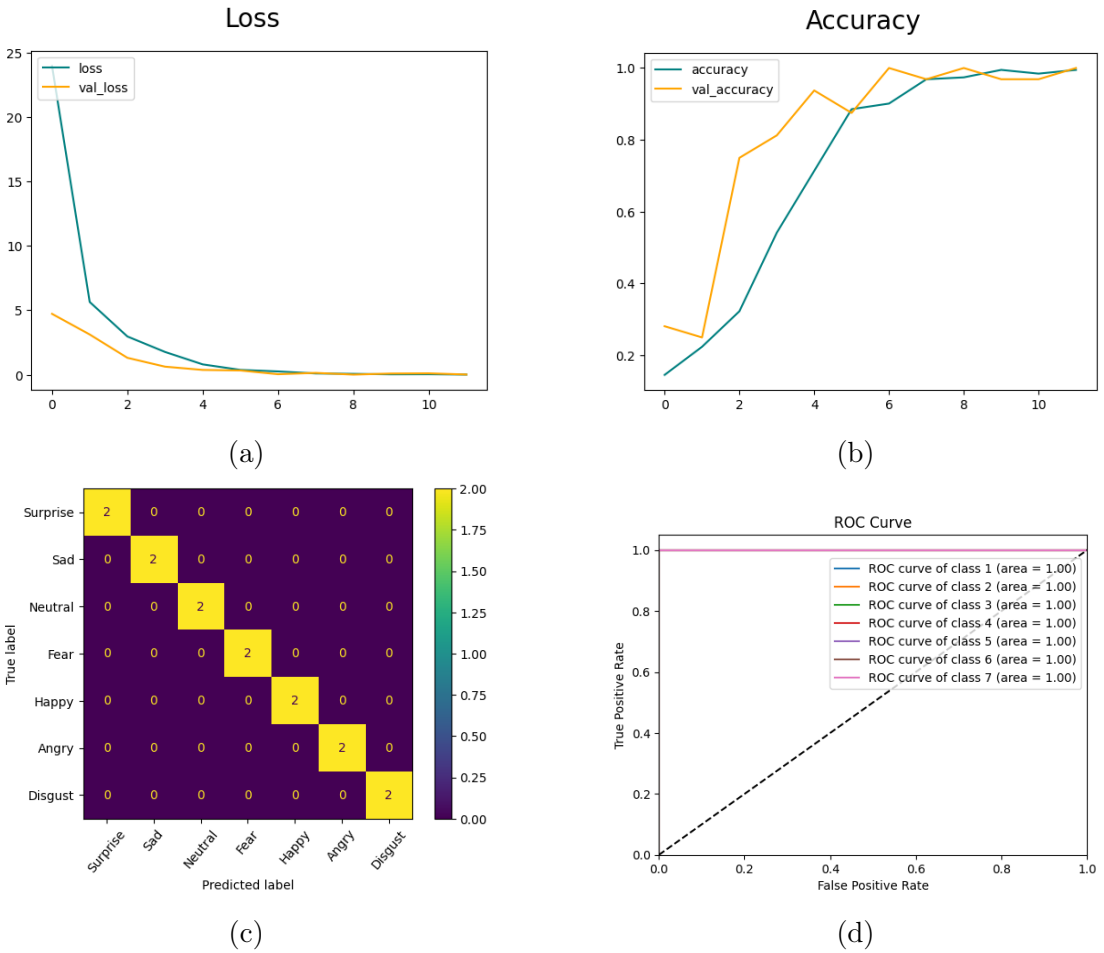


Figure 3.10: Loss Curves, Accuracy Curves, Confusion Matrix and ROC-AUC Curves of the Proposed Model on **Jaffe Dataset**: (a) Loss Curve, (b) Accuracy Curve, (c) Confusion Matrix, (d) ROC-AUC Curve

From the figures in Figure 3.10 We can see that the loss curves suggest that our model is learning well and lowering loss as expected. The accuracy curves show that our model is very accurate and that it doesn't seem to be overfitting very much. The confusion matrix shows how well the model can categorize the test dataset, which is quite well. The final ROC curve for this dataset indicates that the False Positive rate is very low. Moving the graph to the top left corner shows how well the proposed model can sort test images into one of the four classes.

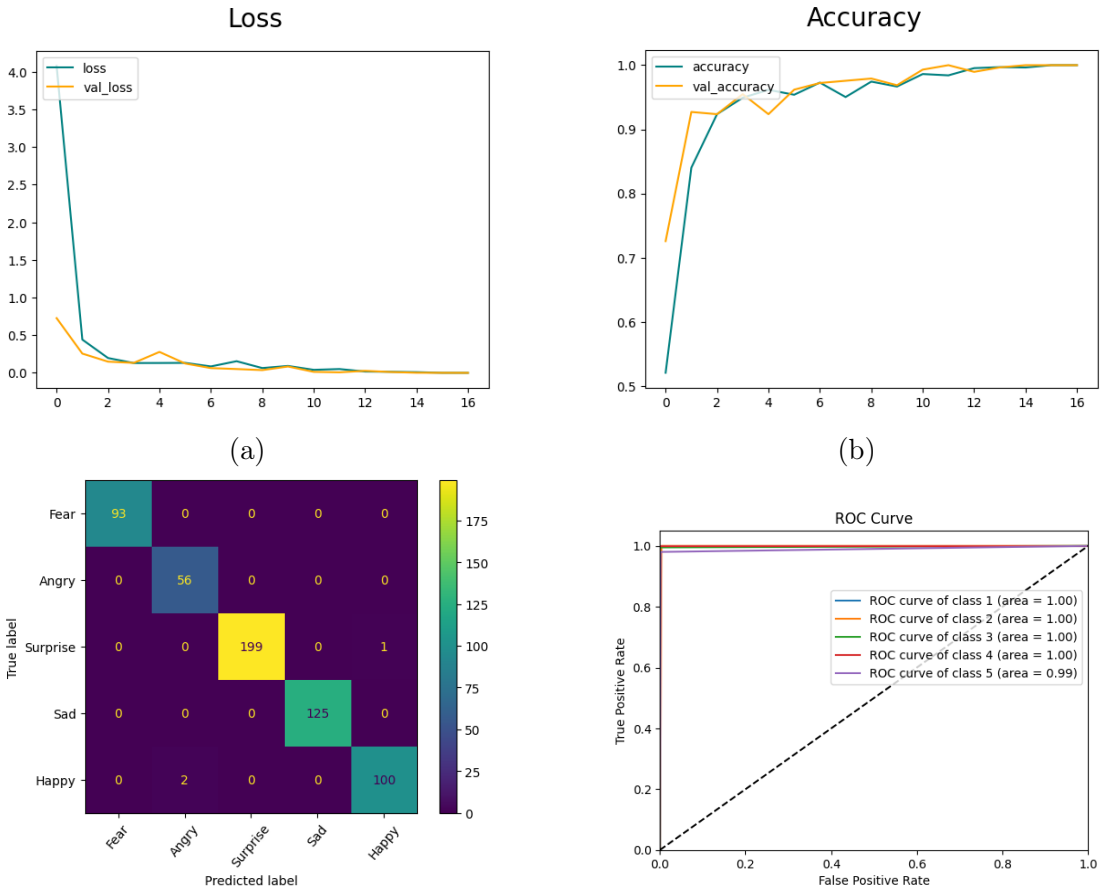


Figure 3.11: Loss Curves, Accuracy Curves, Confusion Matrix and ROC-AUC Curves of the Proposed Model on **Cohn-Kanade Dataset**: (a) Loss Curve, (b) Accuracy Curve, (c) Confusion Matrix, (d) ROC-AUC Curve

Once more, we can see from the figures in Figure 3.11 that our model can learn quickly and lower loss as we intended, as seen by the loss curves. The accuracy curves illustrate how well our model works and that it doesn't seem to have any problems with overfitting. The confusion matrix shows how well the model works at sorting data in the test dataset. The final ROC curve for this dataset shows how very low the False Positive rate is. The graph moves to the upper left corner, which illustrates how well the suggested model can sort test images into different class Table 3.3 provides a detailed ablation study of our proposed approach in following steps: (i) MobileNetV2 only, (ii) MobileNetV3Small only, (iii) MobileNetV3Large only, (iv) (i)+(ii)+(iii)+Swish1, (v) (iv) + Inception blocks i.e., the Proposed model. This in turn highlights the efficiency and contribution of each part of our proposed methodology.

Jaffe Dataset				
Models	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
(i)	97.65	97.57	97.96	97.76
(ii)	97.59	97.60	97.59	97.59
(iii)	98.86	98.83	98.85	98.84
(iv)	99.12	98.97	98.88	98.92
(v)	99.57	99.59	99.42	99.50
(vi)	100.0	100.0	100.0	100.0
Cohn-Kanade Dataset				
(i)	96.88	96.52	96.71	96.61
(ii)	96.92	97.01	96.80	96.90
(iii)	97.21	96.99	96.93	96.96
(iv)	97.67	97.24	97.33	97.28
(v)	98.76	98.69	98.68	98.68
(vi)	99.48	99.11	99.51	99.30

Table 3.3: Ablation Study of our proposed model on the individual dataset

3.5.4 Comparison with state-of-the-art methods

We created our model using the fusion of different MobileNet backbones in the way we advised. When the suggested strategy is compared against methods from the literature, it is shown to have the greatest accuracy score. We have used the Jaffe and Cohn-Kanade (CK+) dataset and compared it to other. Table 3.4 presents an illustration of a comparison between our proposed model and others.

Our tests show that the suggested strategy works the best, with 100.00

Methods	**Approach**	**Accuracy (%)**	
		Jaffe	**CK+**
Mayya et al. 2016 [87]	Caffe ImageNet	98.12	97.00
Kim et al. 2019 [88]	Hierarchial FER Algorithm	91.27	96.46
Bendjillali, et al. 2019 [89]	DWT Feature for Deep CNN	98.43	96.46
Li, K et al. 2020 [90]	CNN with new face cropping and rotation strategy	97.18	97.38
Gonzalez-Lozoya et al. 2020 [91]	VJ Algorithm and CNN	98.26	89.40
Chirra et al. 2021 [72]	DCNN-VC	99.57	99.04
Minaee et al. 2021 [92]	Novel Attention CNN	92.80	98.00
Proposed Model	**An ensemble of CNN models**	**100.00**	**99.48**

Table 3.4: Performance comparison of the proposed model with state-of-the-art methods on the Jaffe and CK+ datasets. Results are in %

3.5.5 Data Visualization

We use GradCAM and t-SNE plots in this part to graphically highlight some of the findings of the suggested method.

3.5.5.1 GradCAM Analysis

This study used GradCAM, a technique that, as noted in [93], produces a gradient-weighted class activation map, to create visual representations of model predictions. These pictures help to make the decision-making process of neural networks clearer. We have utilized GradCAM to make visualizations of the facial photos in the Jaffe and CK+ datasets using the models used in this work, as Figure 3.12 and Figure 3.13 illustrates.It's evident that the different models focus on different parts of the images. This means that different learners get varied and useful information based on their own unique architectures. This also shows that an ensemble technique works well because the different traits that were retrieved can be combined before the final prediction.

3.5.5.2 t-SNE plots

According to [94], t-SNE (t-Distributed Stochastic Neighbor Embedding) is a well-known method for reducing the number of dimensions in data so that it can be displayed in a lower-dimensional space. The program first changes the high-dimensional Euclidean distances between the data points into conditional probability scores that show how similar they are. To do this, we employ SNE (Stochastic Neighbor Embedding) on the data points. The Equation 3.8 is used to define the conditional probability $P_j|i$, which indicates how similar data points x_j and x_i are to one another.

$$P_{i|j} = \frac{exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})} \tag{3.8}$$

We see that the images that show distinct emotions are clearly grouped together in different clusters of points. The first and second images show the t-SNE plot visualizations for the emotion classes of facial emotion photographs from the Jaffe and Cohn-Kanade datasets that the model made.of Figure 3.14 respectively.

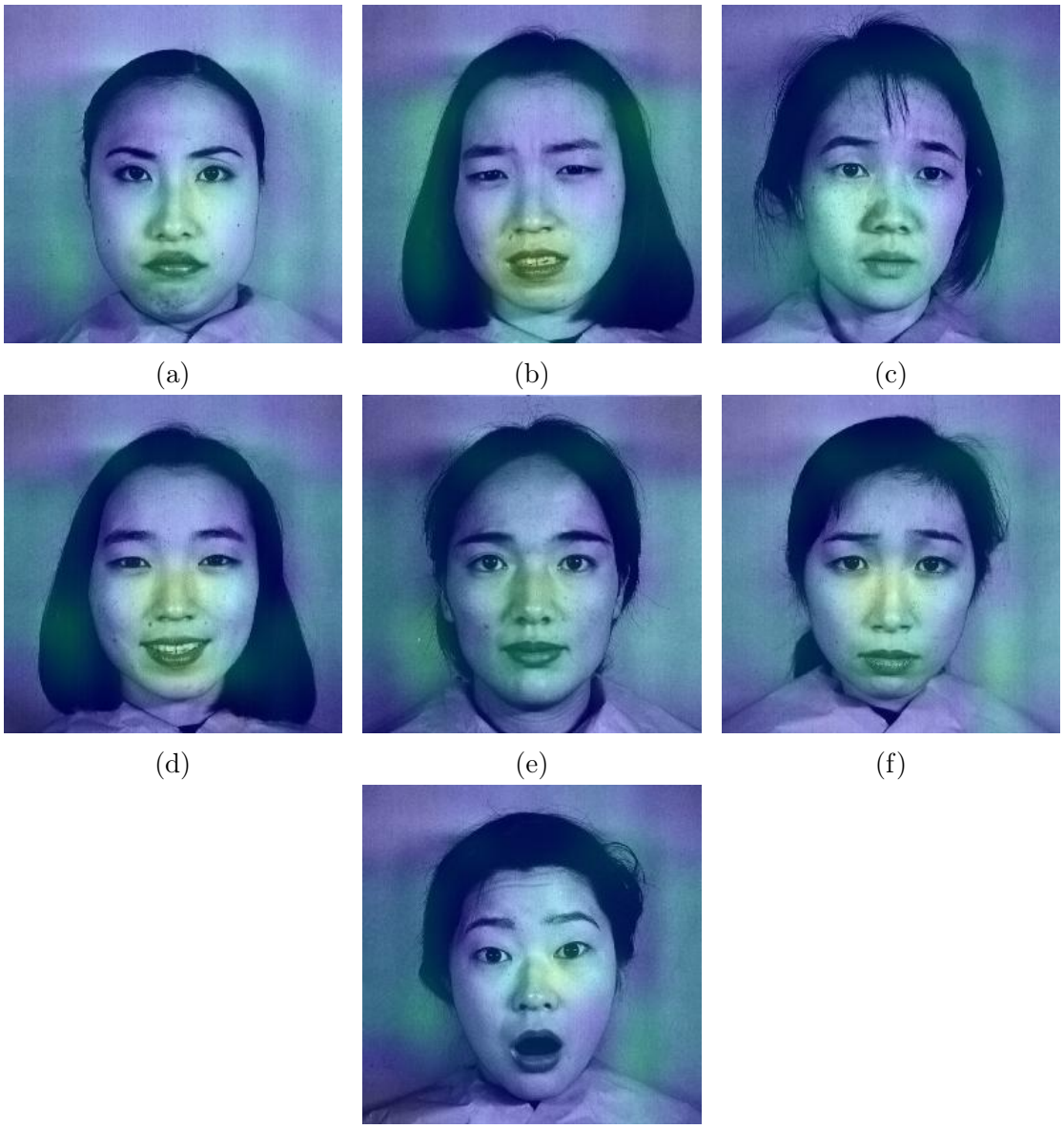
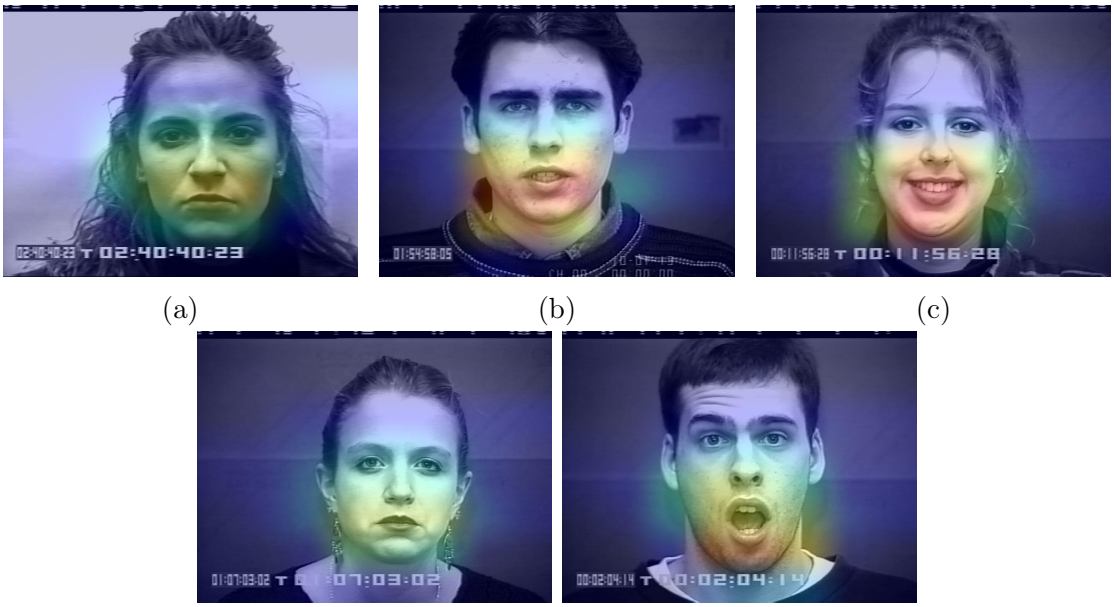


Figure 3.12: GradCAM of our Model on **Jaffe Dataset**: (a) GradCAM of Angry, (b) GradCAM of Disgust, (c) GradCAM of Fear, (d) GradCAM of Happy, (e) GradCAM of Neutral, (f) GradCAM of Sad, (g) GradCAM of Surprise

3.6 Summary

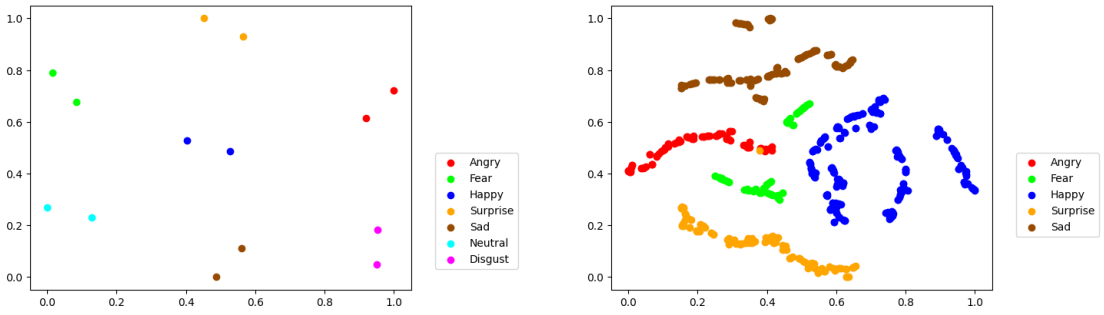
Recent studies indicate that automating the identification process can enhance the precision of emotion recognition and diminish the likelihood of human mistake in computer-aided systems. To enhance the precision of a facial emotion detection model, we have developed an innovative MobileNet fusion model in this study. To do this, we employed the backbones of MobileNetV2, MobileNetV3Small, and MobileNetV3Large. This offers each branch of the model a different level of importance, taking into account how unsure each one is about the predictions. Finally,



(a) (b) (c)

(d) (e)

Figure 3.13: GradCAM of our Model on **Cohn-Kanade Dataset**: (a) GradCAM of Angry, (b) GradCAM of Fear, (c) GradCAM of Happy, (d) GradCAM of Sad, (e) GradCAM of Surprise



(a) (b)

Figure 3.14: t-SNE plots of our Model on **Jaffe and Cohn-Kanade Dataset**: (a) t-SNE plot on **Jaffe Dataset**, (b) t-SNE plot on **Cohn-Kanade Dataset**

the data from each of them are combined to make a final prediction model that is more accurate than the predictions generated by each one on its own. The proposed method is evaluated using the publicly accessible Facial Emotion Recognition datasets, Jaffe and Cohn-Kanade (CK+), and the results exceed those of numerous other previously proposed methodologies. There are, however, false positives and false negatives, which is a big problem for the industry because they affect the detection and identification process right away. So, we need to stop making these mistakes in the future. In the future, we might add some attention mechanisms to the basic CNN models to make stronger feature maps and, in the end, a model that makes better predictions.

We might use similar methodologies, such as that of Yang et al.[95], Jiang

et al.[96], Shenhua et al.[97], Zhu et al.[98], Jebrni et al.[99] and other fusion or multi-modal techniques such as Liu et al.[100] and Ruhan et al.[101]. We may also add some lightweight CNN models in the future to make the system more useful in real-world situations. We want to point out that, even if the performance is limited by the availability of datasets and resources, we plan to test it on more difficult real-world datasets in the future, such as FER2013 . Additionally, the anisotropic spherical Gaussian technique [102] can be utilized to improve the model's resilience to biases in the future. Wang et al. [103] suggested that we could use deep nearest centroid approaches for visual recognition. This could work with our method and make it better.

CHAPTER 4

Modified Sigmoid function-based Ensemble Network(Signet) for recognizing facial emotions

One of the most important areas of study in computer vision and artificial intelligence is facial expression identification. In this work, we introduce a new ensemble-based automated face emotion identification system with a special reparameterization of the Sigmoid activation function. Although sigmoid is frequently used as an activation function in deep learning and machine learning models, its usage in ensemble approaches is unusual. In order to improve face emotion identification, our research intends to present a novel method that capitalizes on Sigmoid's characteristics within the framework of fuzzy rank-based ensemble learning. An important contribution of this work is reparameterizing the Sigmoid function for a fuzzy rank-based ensemble approach, which improves its expressiveness for better suitability in ensemble methods. By integrating multiple base models, our approach effectively harnesses diverse information, resulting in superior facial emotion recognition accuracy. Furthermore, the fusion of complementary data from various sources strengthens the accuracy and robustness of the emotion recognition system. This study addresses the significance of facial emotion recognition, a critical aspect in human-computer interaction, affective computing, and various employs for healthcare, amusement, and security. The method's benefits, such as increased accuracy and generalization across datasets, are demonstrated by experimental assessments using the Jaffe and Cohn-Kanade (CK+) datasets.

4.1 Introduction

Facial emotion recognition, a discipline profoundly entrenched in human-computer interaction [36], [37], affective computing [35], and diverse applications spanning healthcare to security [38], [39], [40], [41], has experienced substantial progress in recent years. The capacity to automatically recognize and interpret human emotions from facial expressions is a crucial component in the advancement of more empathic and responsive AI systems. This is the first step toward the new automated face emotion detection system shown in this paper. It uses a new ensemble learning method and a re-parameterized Sigmoid function to solve some of the problems that have been bothering this field.

The comprehension of human emotions via facial expressions has been a subject of intrigue and investigation since the inception of psychology and cognitive science [42], [43]. The endeavor to elucidate the complexities of human emotional states by facial clues originated from the research of Charles Darwin in the late 19th century. Darwin's groundbreaking study "The Expression of the Emotions in Man and Animals" [44] established the groundwork for the methodical investigation of emotional expression via the human face, a subject that would be thoroughly examined by subsequent scholars, including Paul Ekman [45], [46].

Over the years, advancements in computer vision and machine learning have breathed new life into the pursuit of automatic facial emotion recognition. These advancements in technology allow us to utilize computational tools to decipher the intricacies of human emotions [47], [48], [49], [50]. But the job is not easy because human facial expressions are naturally complicated and changeable. Old methods, which generally relied on hand-made features and single-model classifiers, had a hard time capturing all the subtle emotional differences.

To overcome these challenges, researchers have turned to ensemble learning techniques, which combine multiple models to make predictions, effectively leveraging the diverse information captured by each constituent model [104]. Ensembles offer improved robustness, generalization, and results. The effectiveness of ensemble methods has been demonstrated across various domains of machine learning, including facial emotion recognition [105].

In this context, the current study introduces a novel approach that reimagines the role of the Sigmoid activation function, a long-standing workhorse in machine learning and deep learning, within ensemble learning for facial emotion recognition. While the Sigmoid function has been predominantly used as an activation function [106], [107], its application in ensemble methods is unique and provides a new perspective on improving emotion recognition results.

When we think about how facial expression recognition is used in the real world,

it's clear that these kinds of improvements are needed. Emotion recognition is very important because it improves how people and computers interact and makes it possible to analyze emotions in the entertainment business.. In healthcare, it can assist in mental health diagnostics and treatment monitoring. In security and surveillance, it aids in identifying potential threats or distressed individuals. Therefore, the motivation for advancing the field lies in the potential to create more accurate, versatile, and reliable emotion recognition systems that can benefit society in myriad ways.

Challenges Faced: As we worked on making our new automated system for recognizing facial emotions, we ran into a few problems that made it hard to come up with new ideas. One of the biggest problems we had to solve was coming up with an ensemble method that could bring together all the different models in a useful way. It was hard to coordinate the different pieces of information that these models collected while still keeping the computer running quickly. Also, reparameterizing the Sigmoid activation function for ensemble learning required careful changes and fine-tuning because this new use of the activation function had not been tried before. Ensuring that our strategy works on a variety of datasets and dealing with any overfitting problems were also difficult tasks. These problems, along with others relating to data pretreatment and model selection, were very important in influencing how we came up with our unique method. In the next parts, we'll talk about how we solved these problems and the creative ideas that helped us make a successful facial expression recognition system.

4.2 Literature Review

The literature review in this work is divided into two parts. The first section provides a thorough overview of earlier research on the identification of facial emotions. Finally, the latter part focuses on group methods, in particular, face emotion detection from face images.

4.2.1 A Review of Facial Emotion Recognition

Over the past few years, a number of excellent and useful techniques for facial expression identification have evolved [51],making it one of the most popular areas of research in facial expression recognition. For example, Chen created a novel, unbalanced fuzzy support vector machine.that enhanced the accuracy of expression detection and successfully addressed the issue of data imbalance by adding features for denoising and class compensation[52].

Emotion recognition tasks were initially completed using Deep Belief Network

(DBN) [53], [54], [55]. The face image can be applied directly and doesn't need to be preprocessed by using DBN to parse it.Ranzato et al. employed a gated Markov random field to discern facial expressions [56]. Additionally, DBN employs an extensive corpus of facial images to discern diverse facial expressions. It has also been applied to pre-training data samples, employing a Gabor filter for feature extraction within the deep architecture [57].

To introduce micro-expression algorithms, Ben proposed a system that used dataset alignment and active learning [58]. This essentially solved the problem of not having enough data and made the model better at classifying things. Xu et al. [59] introduced a face expression recognition method utilizing Wasserstein generative adversarial networks.. It creates networks for face emotion and facial identity identification and increases recognition accuracy and robustness through inhibition of intra-class variation.

Utilizing an advanced gravitational search technique based on quantum mechanics and a hybrid deep neural network model, Kumar presents a fusion framework [60]. When addressing local optimum and stochastic characteristics, this approach has significant optimization importance. It was experimentally shown that Zhu et al.'s cascaded face recognition network,which includes temporal feature extraction, hybrid attention, and spatial feature extraction, is better able to handle variations in head deflection, facial posture, and uneven lighting[61].

Li et al. recognized micro-expressions for cross-databases using an unsupervised technique based on distribution adaptation [62]. This technique addresses the issue of a dramatic drop in recognition accuracy brought on by test and training samples from various domains. To minimize feature redundancy and make the network better at learning representative features, Xie proposed a convolutional neural network with a more compact picture representation [63]. Agrawal introduced two novel CNN architectures that improved the model's precision in facial expression recognition and facilitated rapid training through the evaluation of diverse kernel sizes and filter quantities in current CNN models [64].

However, all of the above methods only give a tiny performance improvement because they all employ one model to classify facial emotions.

4.2.2 Survey of Ensemble Techniques

The purpose of ensemble learning is to make the learning framework more accurate and stable by mixing many different learning models. In this section, we will give a short overview of numerous new ensemble learning methods that have been suggested, with a focus on recognizing emotions in faces.

When the ensemble learning technique is employed for facial expression recog-

nition, it often yields higher classification outcomes. For example, Wen did not compute random gradients and made sure that his integrated convolutional echo state network for recognizing facial emotions had a lot of different types of integration [67]. Yu et al. proposed two methods to minimize hinge loss and log-likelihood loss for learning classifier ensemble weights, achieving favorable outcomes on the FER dataset [68]. Pons et al. improved the classifier ensemble by adding supervised learning to the ensemble computation, which raised the FER recognition rate of the ensemble system [69]. Zia et al. created a FER ensemble system with incremental learning capabilities and a dynamic weight majority voting method for base classifiers to make it more accurate and durable over time and space[70].

Deepak et al. [71] proposed an extreme learning machine (ELM) along with ensemble technique using bagging for facial expression recognition. The facial image was split into many little cells in order to get the histogram of orientation gradient (HOG) in their work. Then, numerous distinct bags of training data were created using a bagging algorithm, and each bag was trained with a different ELM. The results were combined using an algorithm that employed a majority vote. The ELM ensemble with bagging makes the network much better at doing general tasks. To assess how well the suggested categorization method performed, two datasets of facial expressions were used: JAFFE and CK+.

Sultan et al. [70] introduced a mechanism known as Dynamic Weight Majority Voting (DWMV) for base classifiers. to maintain elevated precision and resilience of a FER system across spatial and temporal dimensions. The method was made to learn in small increments, so it could learn any expression pattern that could be made in a feature or across multiple ethnic and cultural backgrounds. Venkata et al. [72] proposed a multi-block deep convolutional neural network (DCNN) model to recognize facial expressions in stylized, virtual, and human figures. They created a multi-block DCNN with four blocks, each with its own computational elements, to pull out the features that set facial photographs apart from each other. Two more models were created using ensemble learning to enhance stability and prediction accuracy: the ensemble of three distinct classifiers with a voting algorithm (DCNN-VC) and the bagging ensemble with SVM (DCNN-SVM). Danyang et al. [73] developed graph-based dynamic ensemble pruning (GDEP), an innovative technique for dynamic ensemble pruning, which they applied in the facial expression detection sector. The main purpose of the GDEP is to fix the problem that the classifier selection process is very sensitive to the neighborhood membership of the test sample when utilizing dynamic ensemble pruning approaches. It was used on the CK+ and Jaffe datasets.

The diagram shows the overall flow of the suggested model. in Figure 4.1

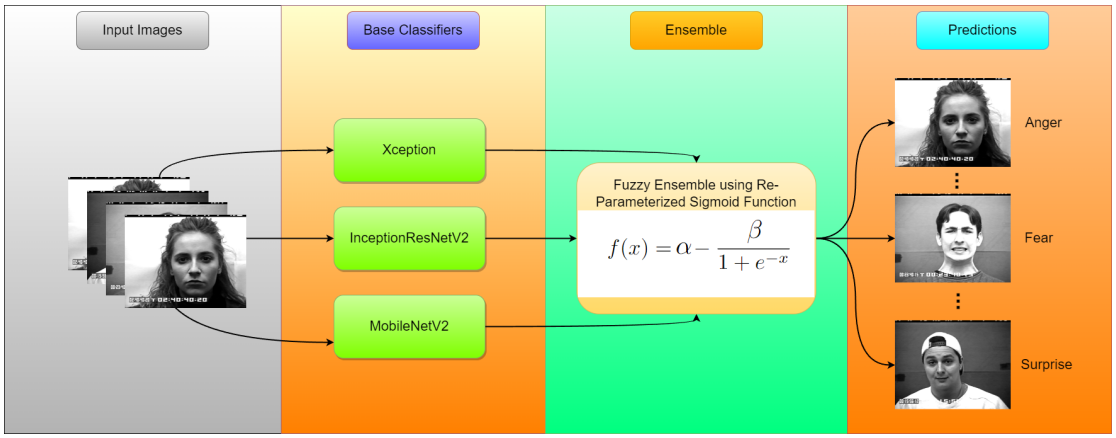


Figure 4.1: Overall pipeline of the proposed model, called SIG-Net, highlighting the base models and the ensemble procedure.

4.3 Dataset Description

The Jaffe and Cohn-Kanade datasets are well-known and regularly utilized in the field of facial expression recognition.

4.3.1 Jaffe Dataset

We may find the Jaffe Facial Emotion Recognition Dataset at https://zenodo.org/records/3451524. It is a very popular benchmark dataset for computer vision and affective computing. The goal of creating it was to help study in face expression identification, which is an important part of understanding how people and computers interact and how to analyze emotions. Michael J. Lyons et al. [74] made the dataset and named it after themselves.

The Jaffe dataset contains pictures of Japanese women's faces, each showing one of seven different facial emotions. Some of these feelings are neutral, happy, sad, surprised, angry, disgusted, and scared. We got the dataset by taking high-resolution, black-and-white pictures of 10 models (5 women) who posed for the camera while acting out the emotions. For each feeling, each model made a set of pictures. The pictures were taken in a controlled setting with consistent lighting and background to reduce variability.

The dataset is split up into subdirectories, one for each of the seven emotions. In each subdirectory, the photographs are given a unique name, like "KA.AN1.39.tiff." The "KA" stands for the model's initials, the "AN" stands for anger, and the number shows which occurrence of that emotion it is.

The Jaffe Facial Emotion Recognition Dataset is mainly used to train and test facial emotion recognition algorithms, machine learning models, and deep learning models [75]. This dataset is often used as a standard for tests by researchers

and developers that work in emotion analysis, facial expression classification, and related disciplines [76], [77], [78].Some images from the Jaffe Dataset in Figure 4.2

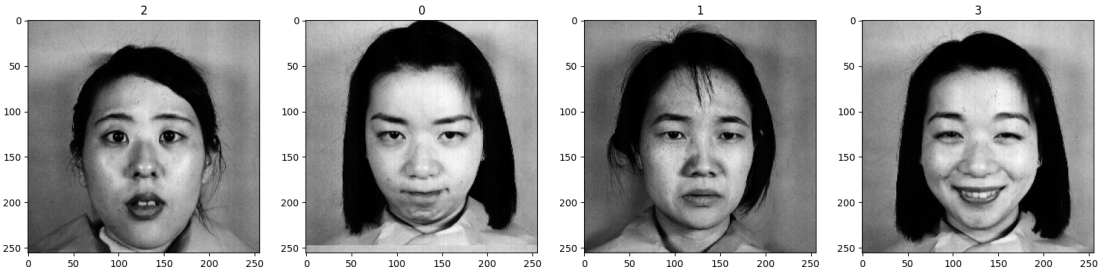


Figure 4.2: Facial emotion images of the Jaffe Dataset.[74][121][161]

4.3.2 Cohn-Kanade (CK+) Dataset

The Cohn-Kanade (CK+) Facial Emotion Recognition Dataset, which can be obtained at `https://www.jeffcohn.net/Resources/`, is a well-known and commonly used benchmark dataset in the fields of computer vision and affective computing. It was made to make it easier to study how to read facial expressions and emotions. Lucey, Patrick and Cohn, Jeffrey F and Kanade, Takeo and Saragih, Jason and Ambadar, Zara and Matthews, Iain [79] are the names of the people who made the dataset.

The CK+ dataset contains a set of high-quality facial photos and emotion labels that go along with them. The photographs show people posing with different emotional expressions. The data for this dataset was collected in a controlled setting, using actors who were told to act out different facial expressions that showed six main emotions and a neutral state. The pictures were taken in well-controlled lighting and on a neutral background to reduce the effects of outside elements on the expressions. The dataset is set up so that each subject has a folder with pictures that show distinct moods. Usually, the pictures have alphanumeric tags on them that tell you what actor and mood they show.

The CK+ dataset is commonly used to train and test algorithms, machine learning models, and deep learning models that can recognize facial expressions. It is the best way to test and compare the performance of different approaches for recognizing emotions. This dataset is used by researchers, developers, and people who work in the domains of computer vision, affective computing, and human-computer interaction for their experiments [80], [81], [82]. Some example pictures from the CK+ Dataset have been givenin Figure 4.3

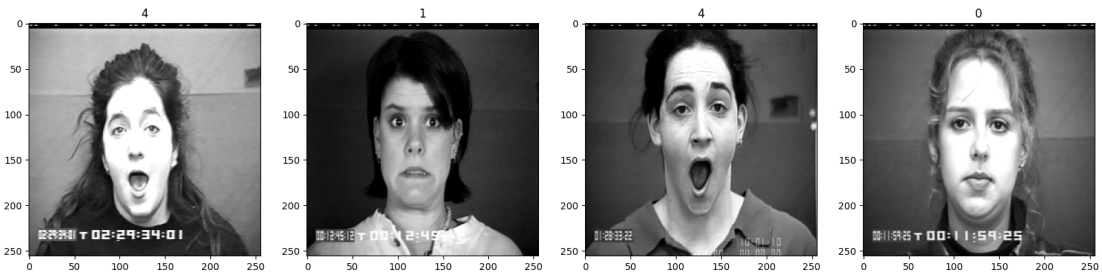


Figure 4.3: Facial expression images of the Cohn-Kanade (CK+) Dataset.[79][122]

4.3.3 Data Preprocessing

The preprocessing we did on the images was to normalize them by dividing them by 255 and then adding weights to the classes to get rid of the class imbalance..

4.4 Proposed Methodology

This section provides a detailed description of the proposed ensemble-based facial emotion recognition model (SIG-Net). We start by providing a short overview of each the base learner that generates confidence scores for an incoming face image in order to identify the proper class of the test data. The suggested ensemble methodology is then utilized to further combine these scores based on rewarding and loss features.

4.4.1 Deep Neural Based Classifiers

The detection of facial emotions from facial image datasets is one of the image classification tasks for which CNNs outperform other types of machine learning systems. One of the main reasons for this is because CNNs are mostly built to work with spatial data, like picture data. Using a combination of convolutional layers and pooling layers, they can successfully capture local patterns and spatial correlations between pixels and learn features from the input data. So, they are really good at figuring out what these kinds of face expression images mean, which often show complex structures and patterns. Also, there is no need for manual feature engineering, CNNs may automatically pick up on and adjust to the features of the incoming data. Overall, these features make CNNs an effective tool for correctly categorizing photos, particularly those of face emotions, and they have the potential to improve identification accuracy and speed. Following a thorough series of studies, we were able to settle on the Xception, InceptionResNetV2, and MobileNetV2 models because they achieved an acceptable degree of accuracy at

the beginning of the entire process.

4.4.1.1 Xception

Chollet et al. [108] presented the Xception convolutional neural network architecture in 2017. It is an extension of the Inception architecture and is referred to as "Extreme Inception" since it replaces the standard Inception convolutions with depthwise separable convolutions. The depth-wise separable convolution layer blocks are the first part of the Xception architecture. After that, batch normalization and ReLU activation are used. The depthwise separable convolution layers are made up of two separate layers. The pointwise convolution layer uses a 1x1 convolutional filter to combine the output channels of the depthwise convolution layer. The depthwise convolution layer applies a single convolutional filter to each input channel. The last layer of the network is made up of a fully linked classification layer and a layer that pools global averages. The depthwise separable convolutions help prevent overfitting and are faster than regular convolutions. The residual connections make the network topologies richer and fix the vanishing gradient problem. So, Xception is an effective and useful model for sorting images.Figure 4.4 illustrates the design of the Xception model.

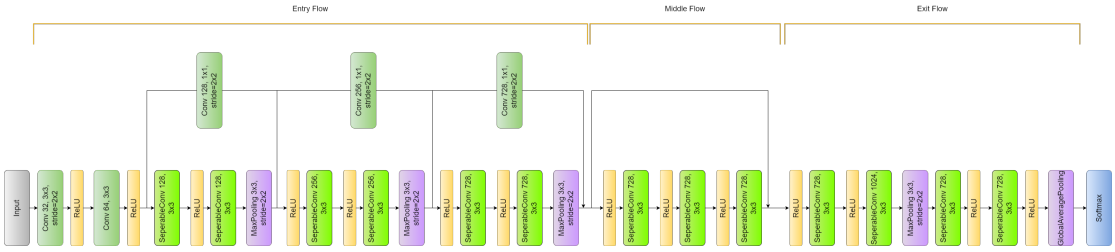


Figure 4.4: Architecture of the Xception model[108]

4.4.1.2 InceptionResNetV2

The benefits of the Inception and ResNet designs are combined in InceptionResNetV2, which Szegedy et al. first presented in 2017 [109]. Inception modules are used in conjunction with residual connections to address the vanishing gradient problem and enable more complex network architectures. In order to achieve high accuracy, the model has fewer parameters and can quickly learn and extract properties from incoming data. The InceptionResNetV2 architecture as illustrated in Figure 4.5.

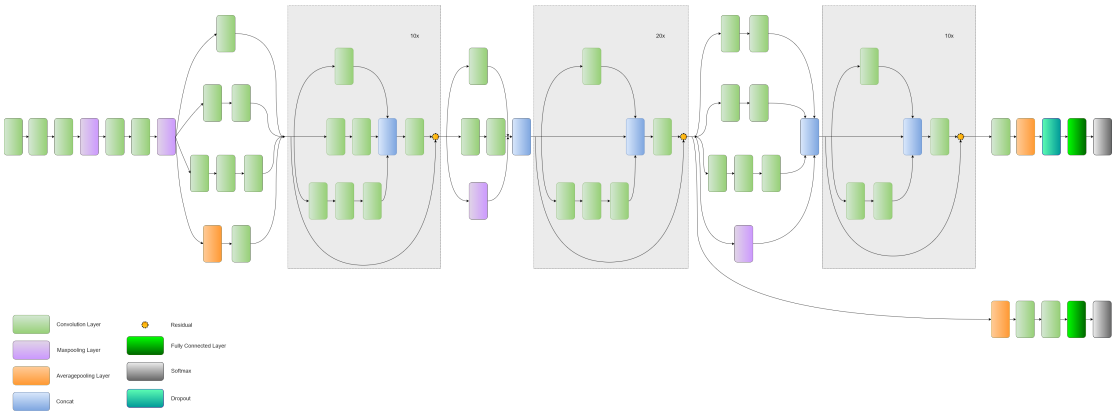


Figure 4.5: Architecture of the InceptionResNetV2 model[109]

4.4.1.3 MobileNetV2

Convolutional Neural Network Sandler et al. made MobileNetV2 in 2018 [19], and it is meant for mobile and embedded vision applications. To make things more efficient while yet keeping their representational capacity, it uses depthwise separable convolution, inverted residuals, and linear bottlenecks. There are two types of blocks: shrinking blocks with a stride of 2 and residual blocks with a stride of 1. There are three levels in each sort of block. The first layer of each block is a 1x1 convolution with ReLU6 activation. The second layer is a depthwise convolution, which uses a distinct convolutional filter for each input channel. The last layer of each block is again a 1x1 convolution, but this time there is no non-linearity. A width multiplier is used in this network to make it work with different hardware and resource limits. MobileNetV2 is an excellent model to utilize when resources are restricted, such on mobile devices, because it works so well and is so light. MobileNetV2 is a small model that works well on several vision tasks, such as semantic segmentation, object detection, and image classification. The architecture of MobileNet is shown in Figure 4.6.

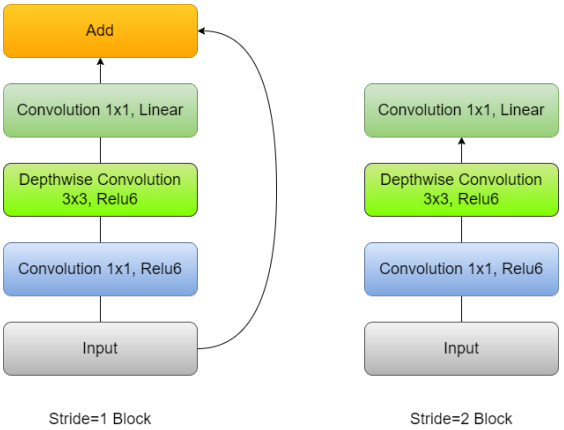


Figure 4.6: Architecture of the MobileNetV2 model[19]

4.4.2 Proposed Fuzzy-Ensemble Approach

The main goal of our suggested method is to give you greater freedom and flexibility while working with datasets that are more or less complicated. We can use a fuzzy ranking-based method to rate each classifier depending on how well it did on a specific test case, taking into account how unpredictable each classifier's predictions are. The proposed methodology employs three CNN-based classifiers (Xception, InceptionResNetV2, and MobileNetV2) to identify facial emotions from images, generating fuzzy scores for each classifier through the re-parameterized Sigmoid function. Then, using this re-parameterized Sigmoid function we used a fuzzy rank based ensemble method; these fuzzy ranks are pooled to improve the classification's overall accuracy.

4.4.3 Significance of Sigmoid function

The Sigmoid function demonstrates how well each model works by plotting a dose response curve between 0 and 1 on the x-axis, which shows how well the model works based on the strength of the input. The Sigmoid function measures the weighted total of the predictions, which tells us how well the forecasts match the actual results. To do this, each model in the ensemble makes a fuzzy rank for a class. Then, the models are put together to make predictions on a validation set for each rank in the top K rankings. The Sigmoid function is used by each base classifier to turn the confidence ratings of a class into fuzzy ranks for that class. You can make an adaptive ensemble model that can adjust the ranking of individual models based on the specifics of each input instance by utilizing this method. Because the decision score of a class that a classifier correctly predicts is often close to one, the re-parameterized Sigmoid function's rapid decline in the range from 0 to 1 helps to create an ensemble of the decision scores from the learning models.

4.4.4 Implementation of the fuzzy ranking ensemble using Sigmoid function

There should to be M distinct decision scores (sometimes referred to as confidence ratings of classifiers), including $CoF^{(1)}, CoF^{(2)}, ..., CoF^{(M)}$ for each input image P. As we utilized three CNN-based models to produce the confidence scores on the dataset, M in our case equals 3(i.e. number of models we have used). In Equation 4.1,the decision scores from the dataset are standardized, with C being the total number of classes in the dataset being analyzed.

$$\sum_{c=1}^{C} CoF_c^{(i)} = 1; \forall i,\ i = 1, 2, 3, ..., M \tag{4.1}$$

Fuzzy ranks are created by using the confidence scores of all the samples in the test dataset, which come from different classes. The re-parameterized Sigmoid function for a class c makes the fuzzy rank using the i_{th} classifier's decision scores as shown Equation 4.2.

$$R_c^{(i)} = \alpha - \frac{\beta}{1 + e^{CoF_c^{(i)}}}; \forall i, c,\ i = 1, 2, 3, ..., M; c = 1, 2, ..., C;\ where\ \alpha = 3, \beta = 4 \tag{4.2}$$

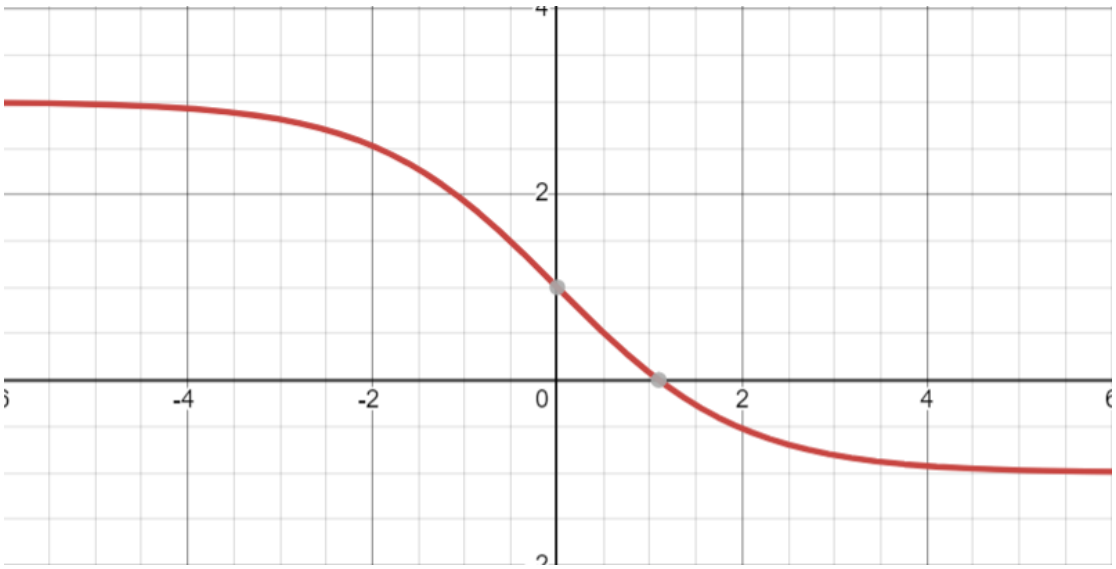


Figure 4.7: Graphical representation of the modified Sigmoid function used in our work.

Figure 4.7 shows the pictorial depiction of the modified Sigmoid function graph, as mentioned in Equation 4.2.

The value of $R_c^{(i)}$ is between 0.0758 and 1, with 0.0758 being the lowest value (best rank). This means that a higher level of confidence leads to a lower (better) value of rank. The fuzzy rank sum (FRS_c) and complement of confidence factor sum ($CCoFS_c$) are calculated in the following way if the top k ranks are shown by $K^{(i)}$, that is, the following are the rankings 1, 2,..., k, that go with class c:[110]:

$$FRS_c = \sum_{i=1}^{M} \begin{cases} R_c^{(i)}, & \text{if } R_c^{(i)} \in K^{(i)} \\ P_c^R, & \text{otherwise} \end{cases} \tag{4.3}$$

$$CCoFS_c = \frac{1}{M}\sum_{i=1}^{M}\begin{cases} CoF_c^{(i)}, & \text{if } R_c^{(i)} \in K^{(i)} \\ P_c^{CoF}, & \text{otherwise} \end{cases} \tag{4.4}$$

If class c does not fall In the top k class ranks, penalties P_c^R and P_c^{CoF} are applied on it. Using the aforementioned Sigmoid function, the value of P_c^R is set to 1, which is determined by setting $CoF_c^{(i)} = 0$; and the value of P_c^{CoF} is set to 0. The penalty values stop class c from becoming an unlikely winner. The decision score in question comes from combining FRS_c and $CCoFS_c$, which is utilized to make the ensemble model's final predictions Equation 4.5 determines the final decision score (FDS).

$$FDS_c = FRS_c * CCoFS_c \tag{4.5}$$

The final projected class for a data instance is the one with the lowest FDS value $\mathbf{I}$, which is provided as shown in Equation 4.6.

$$class(\mathbf{I}) = \underset{c=1,2,...,C}{\arg\min}\, FDS_c \tag{4.6}$$

The overall method of the ensemble approach has been presented in algorithm Algorithm 1.

The entire ensemble process has been further explained with an example in section 4.5.

The value 3 and 4 in Equation 4.2 was determined mathematically as well as experimentally tested. It is determined from the concept of shearing in linear algebra, where a function $f(x)$ needs to be sheared so that it goes through two precise places, $(0, 1)$ and $(1, 0.0758)$. The confidence score of the model can't be more than 1 or less than 0, and the rank has to be within the range that has already been set. The original sigmoid function was inverted as we just needed a negative sloping curve for rank generation due to its inverse relationship. If S was our sigmoid function, then our base function $f(x) = -S$. We can now use shearing to produce the function we want: $g(x) = A * f(x) + B$. We would get $A = 4$ and $B = 3$ by solving for the two points we were given. This means that $g(x)$ is the re-parameterized sigmoid function we utilized in our work.

Algorithm 1 Sigmoid function based Ensemble Method

$C \leftarrow Total\ no.\ of\ classes$
$M \leftarrow No.\ of\ base\ learners$
$p1 \leftarrow Test\ results\ of\ Xception$
$p2 \leftarrow Test\ results\ of\ InceptionResNetV2$
$p3 \leftarrow Test\ results\ of\ MobileNetV2$

function FUZZY_RANK(CF)
    $R\_L \leftarrow 3 - \frac{4}{1+e^{CF}}$ ▷ Applying Sigmoid function to each class of every pixel
    $K\_L \leftarrow Initialise\ array\ of\ R_L\ shape\ with\ penalty\ values$
    if $R\_L_c^i \in top\ 2\ rank$ **then** ▷ R_L for a class results by ith base model for cth class
        $K\_L_c^i \leftarrow R\_L_c^i$
    end if
    return K_L
end function

function CFS_FUNC(CF, K_L)
    if $R\_L_c^i \notin top\ 2\ rank$ **then**
        $CF_c^i \leftarrow 0$ ▷ Applying penalty values
    end if
    $CFR \leftarrow \sum_{c=1}^{C} \sum_{i=1}^{M} CF_c^i$
    $CFS \leftarrow 1 - \frac{CFR}{M}$
    return CFS
end function

function SIGMOID($p1, p2, p3$)
    $CF \leftarrow array(p1, p2, p3)$
    $pred \leftarrow Array\ initialized\ with\ shape\ of\ p1$
    $R\_L \leftarrow$ **fuzzy_rank(CF)**
    $RS \leftarrow \sum_{c=1}^{C} \sum_{i=1}^{M} R\_L_c^i$
    $CFS \leftarrow$ **CFS_func(CF, R_L)**
    $FS \leftarrow RS \times CFS$
    if $FS_c\ is\ minimum$ **then**
        $pred \leftarrow class\ c$ ▷ Predicting the class for each image
    end if
    return $pred$
end function

results ← **Sigmoid(p1, p2, p3)**

4.5 Representation of the entire proposed method

This section provides a detailed schematic representation for the ensemble approach step by step for a single image of CK+ dataset, depicted in Figure 4.8

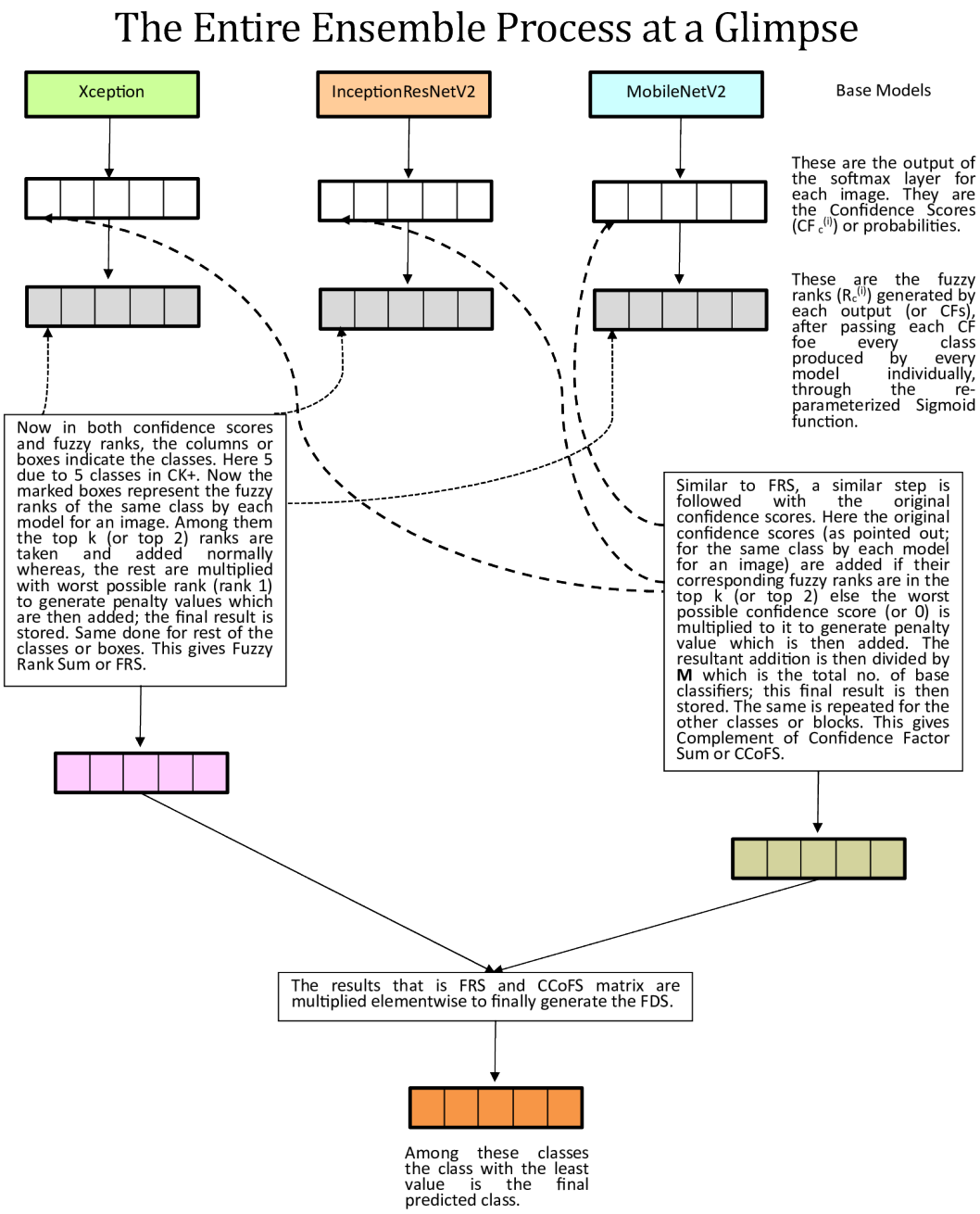


Figure 4.8: Architecture of the proposed model

4.6 Results and Discussions

This section will present the comprehensive results and the associated analysis of the suggested ensemble of CNN models employed for facial emotion recognition from facial photographs. You can already see how the images in the dataset are spread out in section 4.3. Also, the outcomes' effects are talked about.

4.6.1 System Configuration

The full series of experiments was done on a Jupyter Notebook with a 12 GB NVIDIA Tesla T4 GPU that was made available through Google's collaborative environment. We used Tensorflow, Keras, Matplotlib, Scikit, Numpy, and Pandas as the key open-source modules to evaluate our suggested solution in Python.

4.6.2 Evaluation Metrics

Metrics for evaluation are very important for figuring out how effective and powerful a prediction or learning model is. These measures let us figure out how well the model works by measuring how accurately it predicts results. Using a variety of evaluation metrics is important to get a complete picture of how well the model is working and to make sure it meets the requirements of the topic being studied. These measures help us figure out how well a system does at guessing the right class label for an input in classification tasks. Let's look at a case with two classes, one of which is called "positive" and the other "negative."

- **True Positive** (T_P) refers to the number of correctly categorized positive samples.
- **False Positive** (F_P) refers to the number of negative samples that are wrongly labeled as positive.
- **False Negative** (F_N) refers to those samples that are wrongly labeled as negative despite being positive.
- **True Negative** (T_N) refers to the number of correctly categorized negative samples.

For the present work, we use the following performance metrics:

Accuracy: Accuracy is the percentage of occurrences that are accurately predicted over all instances. It is used to see how well a model makes correct predictions. To find it, divide the total number of predictions made by the model by the number of forecasts that were correct. Mathematically, Equation 4.7 represents the accuracy.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4.7}$$

Precision: Precision is the percentage of correct positive predictions out of all positive predictions produced by a model. It is determined by dividing T_P

by the sum of T_P and F_P. Mathematically, precision can be represented as in Equation 4.8.

$$Precision = \frac{T_P}{T_P + F_P} \tag{4.8}$$

Recall: Recall is the percentage of real positive events that the model gets right. It is the ratio of T_P to $T_P + F_N$, as demonstrated in Equation 4.9. A higher recall means that the model is better at finding positive examples. A lower recall means that the model might have missed some positive cases.

$$Recall = \frac{T_P}{T_P + F_N} \tag{4.9}$$

F1 Score: The F1 score is a standard way to measure how well a model works on imbalanced data in a classification job. It combines the model's recall and precision into one number. To find it, add the harmonic means of recall and precision together, as shown in Equation 4.10. A higher F1 Score means better performance.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.10}$$

4.6.3 Implementation

At first, we test a lot of different combinations of CNN models to find the optimum set of base learners for our suggested ensemble technique. The hyperparameters chosen for this experiment are specified in Table 4.1.

Hyperparameter	Value/Name
Optimizer	Adam
Loss function	Sparse Categorical Cross Entropy
Learning rate	0.001
No. of epochs	60

Table 4.1: Hyperparameters of the models

Table 4.2 and Table 4.3 show the accuracies, precisions, recalls and the F1-scores obtained from the different models for facial emotion recognition on the **Jaffe** and **Cohn-Kanade** datasets respectively.

These three models i.e. Xception, InceptionResNetV2, and MobileNetV2 give us an accuracy score of 100.00%, 78.57% and 100.00% respectively on Jaffe dataset and accuracy score of 99.65%, 98.26%, and 97.83% respectively on Cohn-Kanade dataset for classification of emotions. The results show that these models are very reliable for being chosen for this fuzzy ensemble. Their complementary feature

Models	No. of parameters	Jaffe Dataset			
		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Xception	21,387,823	100.00	100.00	100.00	100.00
InceptionResNetV2	54,732,007	78.57	85.71	78.57	79.05
MobileNetV2	2,587,719	100.00	100.00	100.00	100.00
Proposed Ensemble	**78,707,549**	**100.00**	**100.00**	**100.00**	**100.00**

Table 4.2: Performance measure of each model on the Jaffe dataset along with their total number of parameters

Models	No. of parameters	Cohn-Kanade Dataset			
		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Xception	21,387,309	99.65	99.46	99.71	99.59
InceptionResNetV2	54,731,493	98.26	98.07	98.75	98.38
MobileNetV2	2,587,205	98.26	97.83	98.70	98.24
Proposed Ensemble	**78,706,007**	**99.65**	**99.62**	**99.60**	**99.61**

Table 4.3: Performance measure of each model on the Cohn-Kanade dataset along with their total number of parameters

extraction abilities and high classification accuracies make them reliable and suitable candidates for integration into ensemble structure. In our research work the Xception model has 21.38M parameters, the InceptionResNetV2 has 54.73M parameters and the MobileNetV2 has only 2.59M parameters for Jaffed dataset and the models Xception, InceptionResNetV2 and MobileNetV2 have 21.38M, 54.73M and 2.59M parameters respectively also, mentioned in Table 4.2 and Table 4.3. The difference in the number of parameters owns to the fact that in Jaffe dataset there are seven classes so the last dense layer has seven outputs whereas in case of Cohn-Kanade dataset, there are five classes so the last dense layer has five outputs. This discripancy causes the difference in the number of parameters of the last layer which gets reflected in the tables.

The confidence scores for each of the two classes for every image in the dataset are generated by each of these base models as their output. For each individual classifier, these confidence scores are generated and saved for each of the basic models for each image. The three transfer-learned models were each trained separately with the Adam optimizer for 60 epochs (with early stopping) to produce the results as in Table 4.2 and Table 4.3. Figure 4.9 shows the loss curves, accuracy curves, confusion matrices and ROC-AUC curves for Xception, InceptionRestNetV2, and MobileNetV2 on the Jaffe dataset. Figure 4.10 shows the confusion matrix and ROC-AUC Curve after final ensemble using re-parameterised Sigmoid function on the Jaffe dataset.

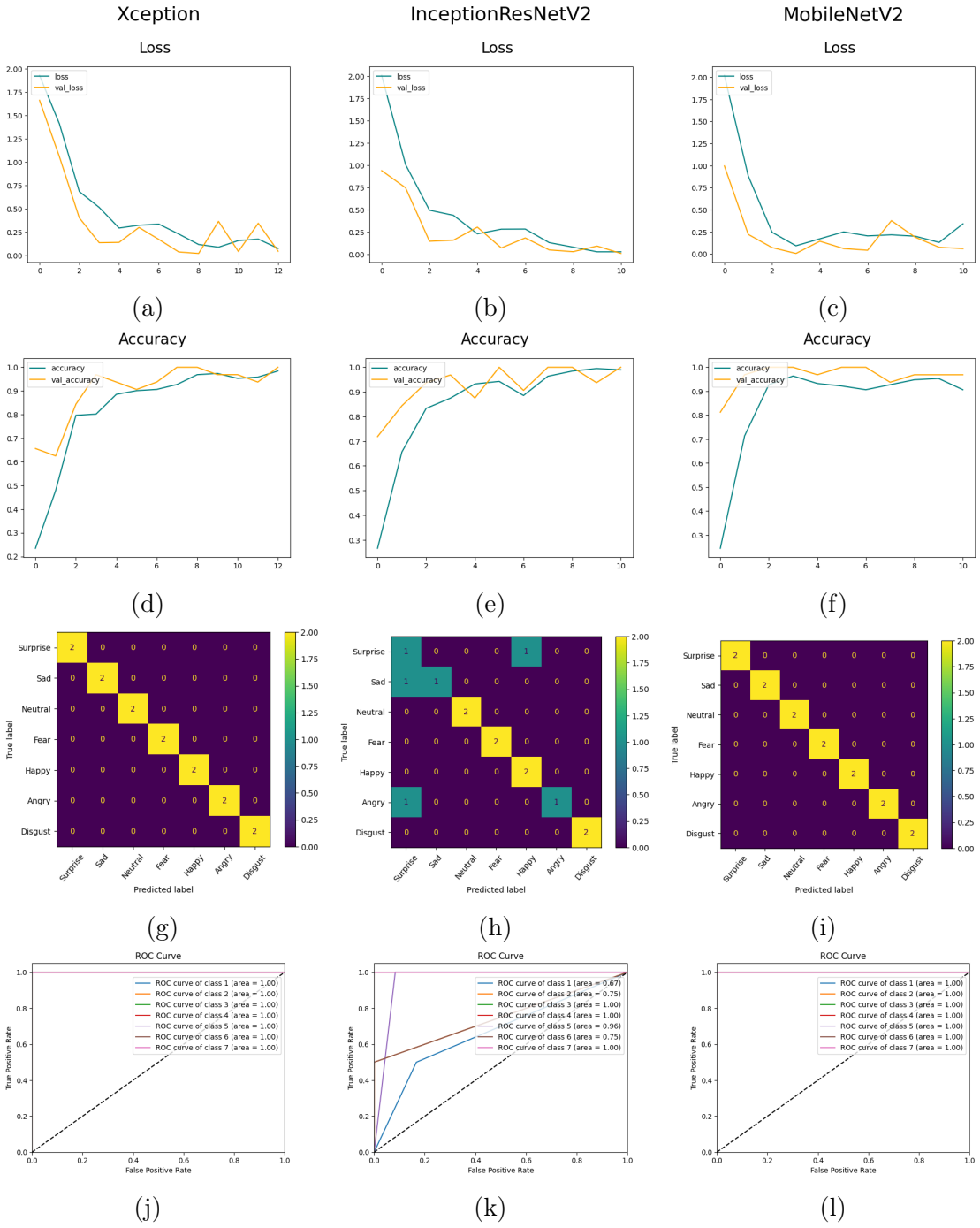
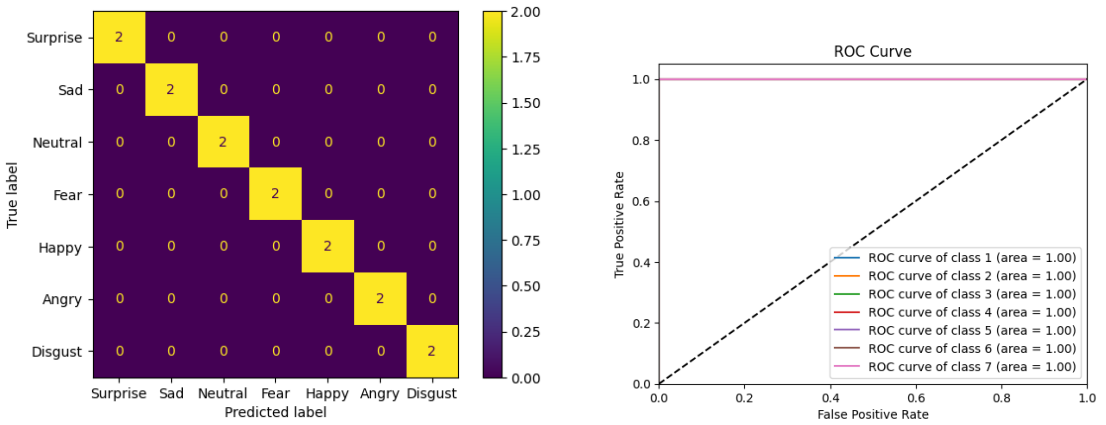


Figure 4.9: Loss Curves, Accuracy Curves, Confusion Matrix and ROC-AUC Curves of the Models on **Jaffe Dataset**: (a) Loss Curve of Xception, (b) Loss Curve of InceptionResNetV2, (c) Loss Curve of MobileNetV2, (d) Accuracy Curve of Xception, (e) Accuracy Curve of InceptionResNetV2, (f) Accuracy Curve of MobileNetV2, (g) Confusion Matrix of Xception, (h) Confusion Matrix of InceptionResNetV2, (i) Confusion Matrix of MobileNetV2, (j) ROC-AUC Curve of Xception, (k) ROC-AUC Curve of InceptionResNetV2, (l) ROC-AUC Curve of MobileNetV2



(a) (b)

Figure 4.10: Proposed Ensemble results using Sigmoid function on **Jaffe Dataset**: (a) Confusion Matrix after Ensemble, (b) ROC-AUC Curve after Ensemble

Figure 4.11 shows the loss curves, accuracy curves, confusion matrices and ROC-AUC curves for Xception, InceptionRestNetV2, and MobileNetV2 on the Cohn-Kanade dataset. Figure 4.12 shows the confusion matrix and ROC-AUC Curve after final ensemble using re-parameterised Sigmoid function on the Cohn-Kanade dataset.

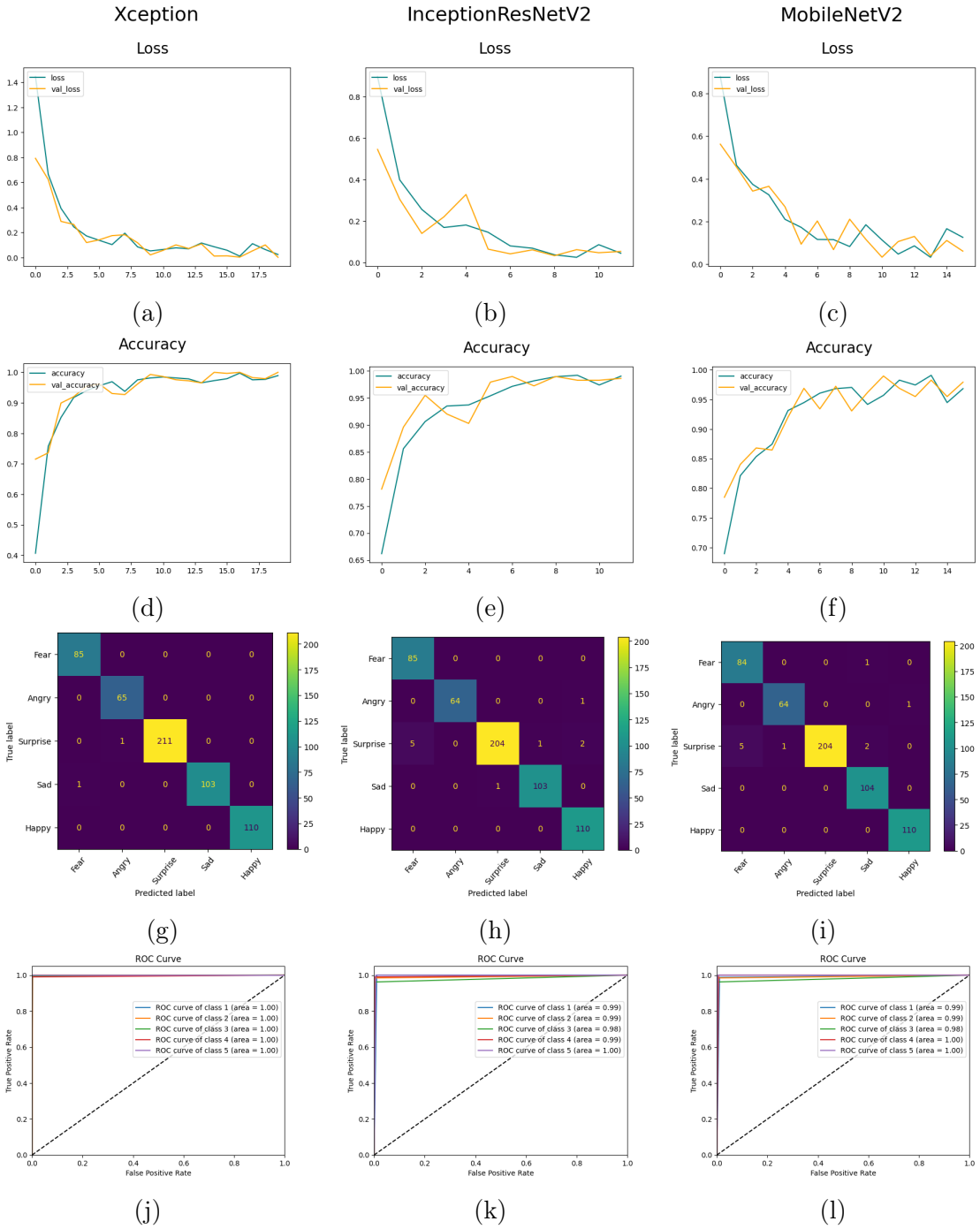
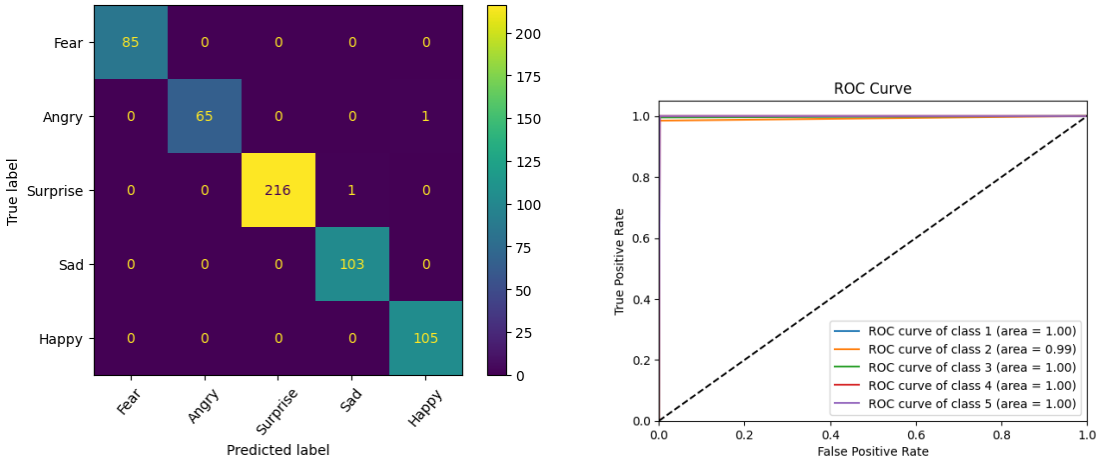


Figure 4.11: Loss Curves, Accuracy Curves, Confusion Matrix and ROC-AUC Curves of the Models on **Cohn-Kanade Dataset**: (a) Loss Curve of Xception, (b) Loss Curve of InceptionResNetV2, (c) Loss Curve of MobileNetV2, (d) Accuracy Curve of Xception, (e) Accuracy Curve of InceptionResNetV2, (f) Accuracy Curve of MobileNetV2, (g) Confusion Matrix of Xception, (h) Confusion Matrix of InceptionResNetV2, (i) Confusion Matrix of MobileNetV2, (j) ROC-AUC Curve of Xception, (k) ROC-AUC Curve of InceptionResNetV2, (l) ROC-AUC Curve of MobileNetV2



(a) (b)

Figure 4.12: Proposed Ensemble results using Sigmoid function on **Cohn-Kanade Dataset**: (a) Confusion Matrix after Ensemble, (b) ROC-AUC Curve after Ensemble

4.6.4 Comparing to the state-of-the-art methods available

We employed the fuzzy ranking-based ensemble method that uses a re-parameterized Sigmoid function to test our model in the way we advised. Fuzzy ranks are made for each base classifier, and then the ensemble method is used to combine these ranks. When the suggested strategy is compared against methods from the literature, it is shown to have the greatest accuracy score. The dataset we have employed and compared with other datasets is the Jaffe and Cohn-Kanade (CK+) dataset. Table 4.4 presents an illustration of a comparison between our proposed model and others.

According to our experiments, the suggested ensemble method provides the highest level of performance accuracy of 100.00% on the Jaffe dataset and 99.65% on the CK+ dataset, when compared to the other methods or classifiers.

Methods	**Approach**	**Accuracy (%)**	
		Jaffe	**CK+**
Mayya et al. 2016 [87]	Caffe ImageNet	98.12	97.00
Kim et al. 2019 [88]	Hierarchial FER Algorithm	91.27	96.46
Bendjillali, et al. 2019 [89]	DWT Feature for Deep CNN	98.43	96.46
Li, K et al. 2020 [90]	CNN with new face cropping and rotation strategy	97.18	97.38
Gonzalez-Lozoya et al. 2020 [91]	VJ Algorithm and CNN	98.26	89.40
Chirra et al. 2021 [72]	DCNN-VC	99.57	99.04
Minaee et al. 2021 [92]	Novel Attention CNN	92.80	98.00
Proposed (SIG-Net)	**An ensemble of CNN models**	**100.00**	**99.65**

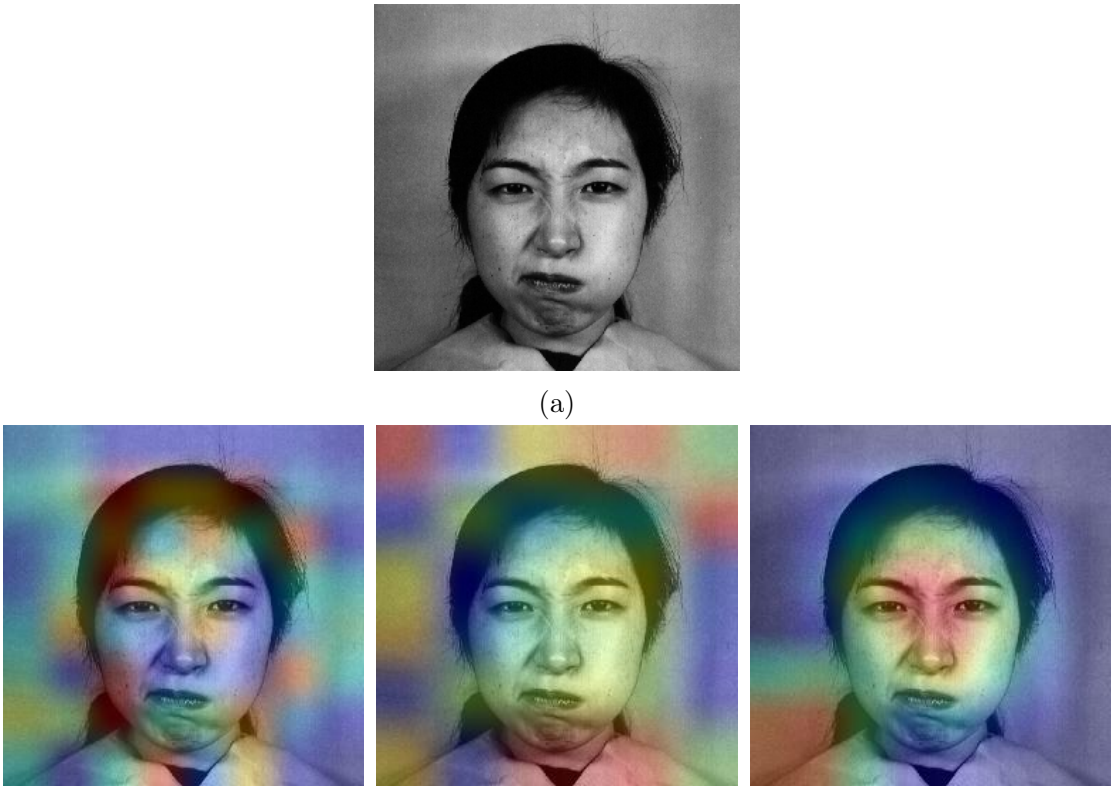
Table 4.4: Performance comparison of the proposed ensemble model with state-of-the-art methods on the Jaffe and CK+ datasets. Results are in %

4.6.5 Data Visualization

In this section, we take help from two data visualization tools to visually show some results of the proposed method namely GradCAM and t-SNE plots.

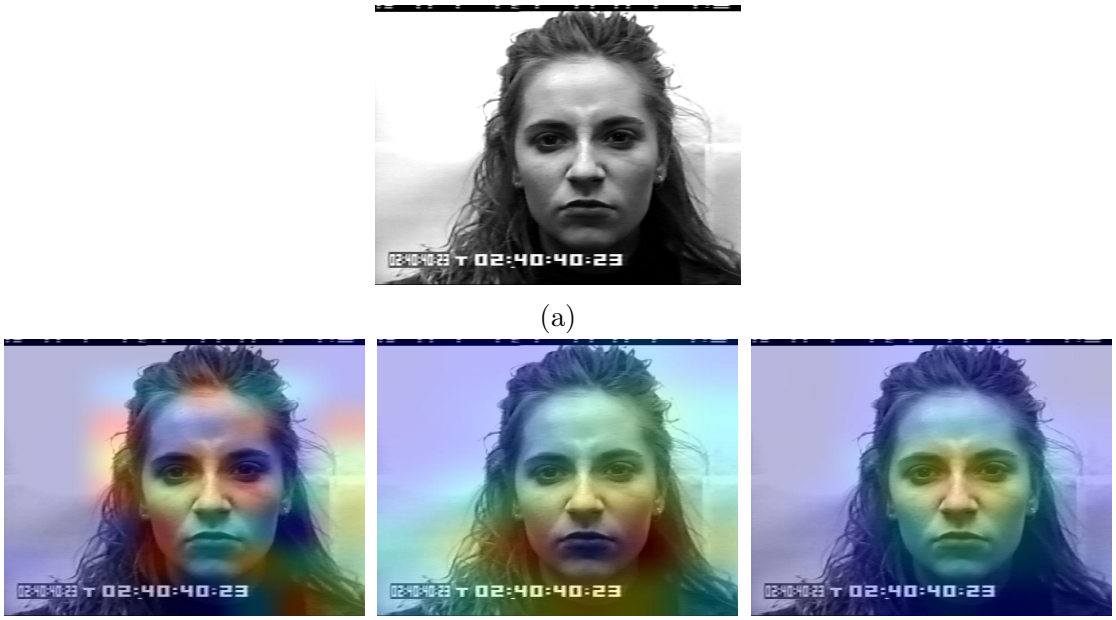
4.6.5.1 GradCAM Analysis

This study used GradCAM, a technique that, as noted in [93], produces a gradient-weighted class activation map, to create visual representations of model predictions. These pictures help to make the decision-making process of neural networks clearer. We have utilized GradCAM to make visualizations of the facial imagess in the Jaffe and CK+ datasets using the models used in this work, as Figure 4.13 and Figure 4.14 illustrates. It's evident that the different models focus on different parts of the images. This means that different learners get varied and useful information based on their own unique architectures. This also shows that an ensemble technique works well because the different traits that were retrieved can be combined before the final prediction.



(a) (b) (c) (d)

Figure 4.13: GradCAM of the Base Models on **Jaffe Dataset**: (a) Original Image, (b) GradCAM of Xception, (c) GradCAM of InceptionResNetV2, (d) GradCAM of MobileNetV2



(a)

(b) (c) (d)

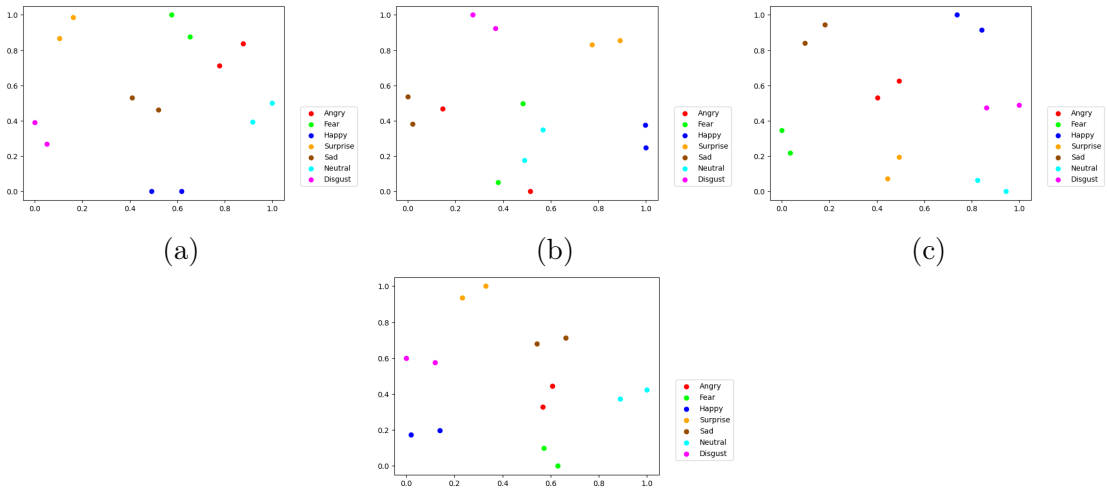
Figure 4.14: GradCAM of the Base Models on **Cohn-Kanade Dataset**: (a) Original Image, (b) GradCAM of Xception, (c) GradCAM of InceptionResNetV2, (d) GradCAM of MobileNetV2

4.6.5.2 t-SNE plots

According to [94], t-SNE (t-Distributed Stochastic Neighbor Embedding) is a well-known method for reducing the number of dimensions in data so that it can be displayed in a lower-dimensional space. The process first changes the high-dimensional Euclidean distances between the data points into conditional probability scores that show how similar the points are. This is done by using SNE (Stochastic Neighbor Embedding) on the data points. The Equation 4.11 is used to define the conditional probability $P_j|i$, which indicates how similar data points x_j and x_i are to one another.

$$P_{i|j} = \frac{exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2})} \tag{4.11}$$

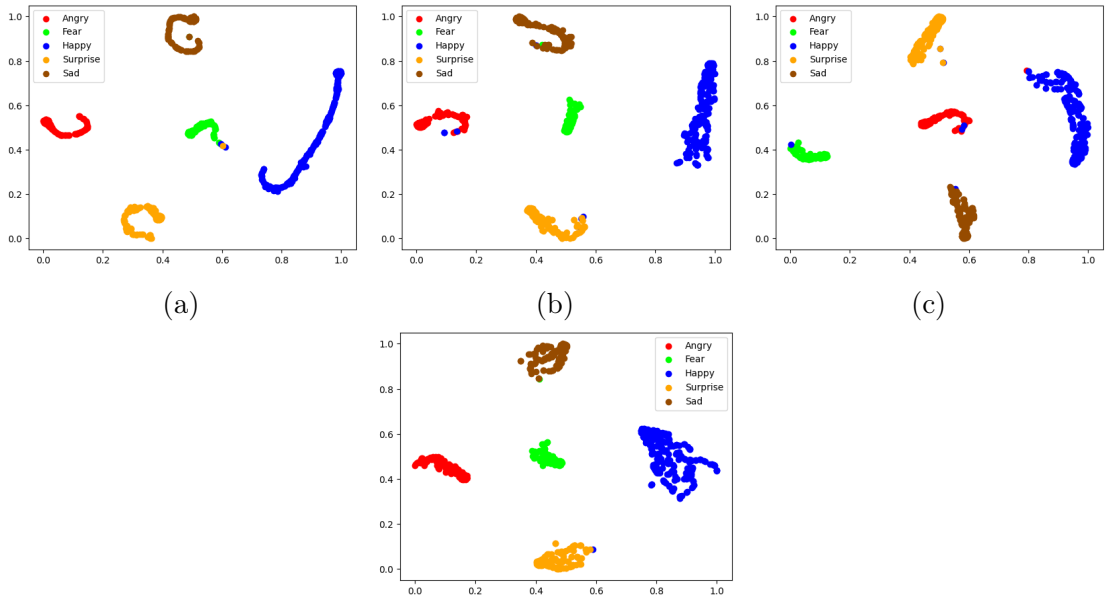
We find that the images related to the different emotions are clearly split into discrete clusters of points. The t-SNE plot visualizations for the emotion classes of facial expression photos from the Jaffe and Cohn-Kanade datasets by the base classifiers are provided in the first three images of Figure 4.15 and Figure 4.16 respectively. The t-SNE plot of the ensemble model is presented in the fourth image of Figure 4.15 and Figure 4.16 respectively.



(a) (b) (c)

(d)

Figure 4.15: t-SNE plots of the Models and Final Ensemble on **Jaffe Dataset**: (a) t-SNE plot of Xception, (b) t-SNE plot of InceptionResNetV2, (c) t-SNE plot of MobileNetV2, (d) t-SNE plot of Proposed Ensemble



(a) (b) (c)

(d)

Figure 4.16: t-SNE plots of the Models and Final Ensemble on **Cohn-Kanade Dataset**: (a) t-SNE plot of Xception, (b) t-SNE plot of InceptionResNetV2, (c) t-SNE plot of MobileNetV2, (d) t-SNE plot of Proposed Ensemble

4.7 Summary

Recent studies indicate that automating the identification process enables computer-aided systems to enhance the precision of emotion recognition and diminish the likelihood of human error. To enhance the precision of a facial emotion recognition model, we have developed an ensemble model in this study called

SIG-Net, which integrates three transfer learning-based CNN models: Xception, InceptionResNetV2, and MobileNetV2. To do this, we used a fuzzy ranking-based strategy that gives each classifier a different level of importance while also taking into account how uncertain the predictions are for each one. The fuzzy ranking system uses the Sigmoid function to make a final prediction model that is more accurate than the predictions provided by each basic classifier on its own. The proposed method is evaluated using the publicly accessible Facial Emotion Recognition datasets, Jaffe and Cohn-Kanade (CK+), and the results exceed those of other recently proposed methodologies. There are, however, false positives and false negatives, which is a big problem for the industry because they affect the detection and identification process right away. So, we need to stop making these mistakes in the future. In the future, we may add certain attention mechanisms to the basic CNN models to make the feature maps stronger and, in the end, the prediction model more accurate. We may eventually add some lightweight CNN models to the system to make it more useful in the real world.

CHAPTER 5

Feature Selection contour for FER with entropy information gain and KNN classifier

Facial expression recognition (FER) is a significantly important task in the extended vision community, having extensive applications in human-computer interaction, and more so in generative vision models. Moreover, its application majorly lies in lightweight devices such as smartphones, CCTV cameras, robots etc. and thus requires models that can deployed on low compute. The challenge, hence, lies in learning an effective feature representation that is both robust and computationally efficient. In this work, we propose a two-stage pipeline for FER, that harnesses the power of deep transfer learning and statistical feature selection to achieve the aforementioned goals. Specifically, in the first stage we adopt a pre-trained EfficientNet model and fine-tune it on our target dataset, followed by extracting high-dimensional features using the frozen backbone. Once extracted, we leverage information gain, a statistical measure based on entropy differences, to quantify the "usefulness" of each of the features and rank them. Taking the top-k subset from the ranked features, we train a k-nearest neighbour classifier to perform the final classification. Upon evaluation, our method proves to be highly competitive, outperforming several existing state-of-the-art works by significant margins on three commonly used FER datasets.

5.1 Introduction

Recognising emotions from different forms of human expressivity, such as speech, face and body language, is an integral element of effective communication.

Figure 5.1: Some samples of facial expression data from each of the datasets used in this work.[123]

While this might seem fairly trivial for humans in general, this has gained a lot of significance in the technical community due to its high impact in human-computer interaction [111]–[113], speech synthesis [114], talking face video generation [115]–[117], among others. In particular, facial expression recognition [118], [119] has earned considerable attention from the vision community, owing to the intrinsic challenges it offers in terms of data and the requirement for fine-grained modelling to identify minute facial expression elements.

Expression recognition can be modelled as a classification problem where each sample belongs to some specific pre-defined emotion class. In particular, for facial expression recognition, the samples comprise cropped human faces depicting a certain expression (anger, sadness, happy etc). The first instinct would be to extract handcrafted features from these faces and train a machine learning classifier, such as support vector machines (SVM) [120], using those extracted features. However, it is noteworthy to observe that though one can obtain facial images according to different emotions, most of them would look similar at a high level as they are all comprising faces (which have a lot of semantic similarity among them). This can be better understood from Figure 5.1 where some facial expression samples have been shown from different FER datasets [121]–[123]. Thus, the primary goal for building a robust FER model would be to recognise intricate fine-grained patterns underlying each emotion class. While traditional ML classifiers fail to do so, convolutional neural network (CNN) [15], [16], [20] models are a possible alternative, which automatically learns translationally equivariant visual features by changing parameter values based on optimization over a loss objective. Due to their automated learning paradigm, they do not require any handcrafted features, which makes the learning process much simpler, efficient and robust.

A significant bottleneck for training a facial expression deep learning model is

the unavailability of large-scale datasets, something that is often widely available for natural images. To the best of our knowledge, there exists only one publicly available dataset – the FER2013 database [123] – that can be considered to be somewhat large-scale, comprising 35k images in total. Other datasets, even the ones used in this study [121], [122] are quite small, less than 1000 images in each. This is in stark contrast with natural images where large datasets such as ImageNet [124] and MSCOCO [125] are available that are abundant in data. In fact, recent advances in the medical imaging community has also led to curation of large-scale radiology [126] and histopathological [127] datasets. This has led to researchers posing manually designed CNN architectures and training them from scratch on these datasets. Although it has achieved promising results, it has two limitations: it requires a lot of manual tuning and designing, and it does not exploit the availability of previously trained state-of-the-art models proposed in literature [15], [16], [20], [128]. A possible solution to this is transfer learning: which "transfers" a previously learned representation (often on a large-scale dataset such as ImageNet [124]) onto the target domain dataset (FER [123] in this case) by fine-tuning on it. This offers two advantages: we no longer need to manually construct the CNN since it is derived from an existing model; and it offers a better initialization point compared to random initialization since the model already has some high-level learned visual features that are generic to several vision tasks [16], [20]. Transfer learning has proven its efficacy in computer vision for non-natural images, particularly medical imaging [129], [130]. In this work, we take inspiration from this paradigm to propose a transfer learning-based feature selection pipeline for facial expression recognition.

Since deep learning models learn a high dimensional feature representation, it often leads to redundancies among them, which not only makes it inefficient (especially for storing them in a database) but also hinders their performance. To mitigate this, researchers have resorted to feature selection [130]–[133] which aims at selecting a subset of the total feature space that is most informative and can represent the overall feature space without its redundancies. There have been efforts in designing feature selection methods based on statistical techniques [131], [133], [134], since they are extremely lightweight and computationally cheap to implement, yet offering high performance. In this research, we adopt Information Gain [134] as a means of statistical measure for feature selection (explained in detail in later sections). There has not been much of research in combining deep feature learning with feature selection for development of a lightweight and efficient pipeline for FER task, which serves as a driving inspiration behind our work.

FER is a task well-suited for computationally lightweight real-time devices (smartphones, CCTV cameras etc), which requires models to be parameter effi-

cient and be able to run on low compute. In this paper, we motivate ourselves to design a lightweight yet highly robust FER framework that can be easily deployed in real-time edge devices. To this end, we propose a two-stage pipeline comprising CNN-based deep learning on raw facial expression images, based on which high-dimensional features are extracted, whose dimensionality is then reduced by leveraging a statistical feature selection strategy. Exploiting the paradigm of transfer learning, we leverage the recently proposed lightweight EfficientNet model [20] pre-trained on ImageNet [124] for feature learning from facial images, after which we extract features from those images using the frozen network. Next, we use Information Gain [134], [135], a filter FS method that ranks features by evaluating the information gain of each feature with respect to the target labels. Based on this ranking, the top-k subset (k=50% chosen) is considered for the final classification task using a nearest neighbours classifier [136]. We validated our approach using 5-fold cross-validation scheme on publicly available FER datasets [121]–[123], where it was found to perform competitively to state-of-the-art approaches [92], [119], [131], [132], outperforming several of them.

In a nutshell, the contributions of this research are as follows:

1. We present a lightweight two-stage deep learning-based framework for recognising expressions from human faces.

2. We combine deep feature learning with the paradigm of statistical feature selection in our framework in order to mitigate possible redundancies in the extracted deep feature set, thus reducing memory requirements as well as improving emotion classification performance.

3. Our model components are suitably chosen so as to be lightweight yet robust, keeping in mind the application of FER models in edge devices such as CCTVs, robots etc.

4. Quantitative evaluation reveals that our proposed framework outperforms several prior arts across three standard FER datasets.

5.2 Related Work

Facial emotion is one of the primary modes of non-verbal communication in humans. With rapid development in Human-Computer Interaction (HCI) and other artificial intelligence systems, the interest and consequent research in automatic Facial Emotion Recognition (FER) is increasing. To leverage the necessity of human FER in industries like robotics, medical sciences, law and driving assistance systems, many FER tools and systems have been developed over the last decade.

Ekman et al. [137]. have defined seven fundamental emotions, which remain consistent across cultures, accompanied by the corresponding facial expressions: anger, fear, happiness, sadness, contempt, disgust, and surprise. In a recent examination of the Facial Recognition Technology (FERET) dataset, Sajid et al. uncovered the significance of facial asymmetry as an indicator for age estimation [138]. Their discovery highlights that right facial asymmetry surpasses left facial asymmetry in accuracy. Detecting facial pose remains a persistent challenge in face recognition, but Ratyal et al. have offered a solution to address the variability in facial pose appearance by using three-dimensional pose invariant approach using subject-specific descriptors [139].

Various conventional approaches have been investigated for automatic FER systems. They typically have the common goal of identifying the facial region and extracting geometric features, or appearance features, or a combination of both from the target face. Ghimire and Lee [120] incorporated two distinct sets of geometric features derived from 52 facial landmarks, calculating angles and distances between landmark pairs, which were then compared to the initial frame. They introduced two classification approaches: one utilizing multi-class AdaBoost with dynamic time warping, and the other applying an SVM to the enhanced feature vectors. Happy et al. [140] employed a real-time facial expression recognition method using LBP histograms of different block sizes from a global face region as feature vectors, classified with PCA, but achieved inadequate results due to a lack of representation for local facial variations in the feature vector. This was improved by Ghimire et al. [141] in which they leveraged region specific appearance features of the face through incremental search approach, thus increasing recognition accuracy. Benitez et al. and Ghimire et al. have successfully combined geometric and appearance features to counter the weaknesses of the two approaches and provide even better results in some cases [142] [141].

While traditional facial recognition techniques relying on manually crafted features have achieved considerable success, researchers have gradually shifted towards deep learning methods in the past few years, attracted by their remarkable automatic recognition capabilities. Lopes et al. [143] proposed a simple way for facial expression recognition by combining Convolutional Neural Network and some image pre-processing methods like data augmentation, rotation correction, cropping, down sampling. Cai et al. [144] introduced a novel CNN architecture with Sparse Batch Normalization (SBP) to address gradient issues using dropout to combat overfitting and tested their approach on various datasets. Ding et al. introduced a novel deep face recognition architecture in FaceNet2ExpNet for FER tasks in [145], which was later improved by Li et al. through transfer learning [90]. Transfer learning approaches were also experimented by Shaees et al. in

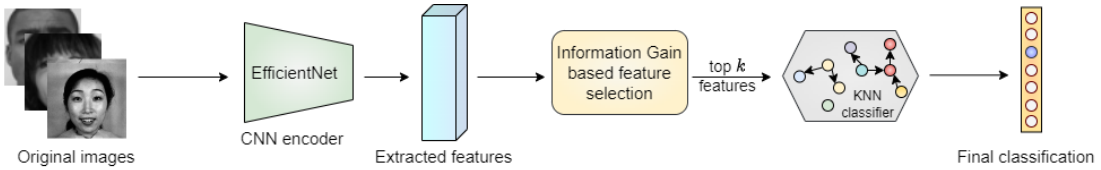


Figure 5.2: A schematic depiction of the proposed methodology for facial expression recognition in this study.

[146]. The authors delved into a hybrid architecture with transfer learning, employing pre-trained AlexNet to extract features that were subsequently classified using support vector machines (SVM). To optimize the dimensional complexities of feature extraction tasks from large datasets, researchers have undertaken feature selection(FS) algorithms. These algorithms aim to reduce the feature redundancy to achieve better overall prediciton accuracy [131] [147]. Mistry et al. introduced a system for recognizing facial expressions, which utilized hvnLBP-based feature extraction and employed a combination of micro GA and particle swarm optimization (PSO) for feature optimization [148]. Filter-based FS methods have been more optimal computationally compared to other metaheuristic approaches to FS tasks in FER like in [149]. The authors implemented a cosine similarity and maximal redundancy-based minimal-redundancy supervised harmony search algorithm, which outperformed benchmark FS methods.

5.3 Proposed Method

As discussed previously, our work is composed of (a) fine-tuning a CNN model on facial expression samples for feature extraction; (b) ranking the features based on a statistical measure known as information gain followed by selecting a top-ranked subset; (c) training a classifier using the top-ranked feature subset for the final classification. The method has been schematically described in Figure 5.2. We now describe these steps in detail in the following subsections.

5.3.1 Feature Extraction using Fine-tuned EfficientNet

The first step in our proposed pipeline is to fine-tune a pre-trained CNN model on our facial expression dataset in order to generate robust feature representations crucial for better classification performance. Keeping in mind the need to develop lightweight methods that can be deployed in low-resource settings (e.g. CCTV cameras, edge devices), we opt for the EfficientNet model [20], a recently proposed state-of-the-art CNN architecture, which is highly parameter efficient as compared to contemporary models like VGG [15], ResNet [16] or DenseNet [150].

Table 5.1: Comparing number of parameters and FLOPs (floating point operations per second) across different state-of-the-art CNN models against EfficientNet. For both, lower value means the model is more efficient. The values have been taken from [20].

Model	**No. of parameters**	**FLOPs**
ResNet-18 [16]	11M	1.8B
ResNet-50 [16]	26M	4.1B
Inception-V3 [128]	24M	5.7B
DenseNet-169 [150]	14M	3.5B
Xception [108]	23M	8.4B
EfficientNet-B0 [20]	**5.3M**	**0.39B**

For context, a table comparing the number of parameters of these models has been shown in Table 5.1.

The basic EfficientNet-B0 model has been used in our work, which has the least number of parameters among all its variants. Following transfer learning paradigm, we load ImageNet [124] pre-trained weights in the network and replace the classification head with a new fully connected layer having the number of classes in the dataset as its output dimension. Specifically, the CNN backbone outputs a feature vector of dimension 1280, which is then fed into the classification layer to get the softmax predictions. We fine-tune the entire network end-to-end using the categorical cross-entropy loss, which is the most commonly used classification objective in deep learning literature.

Once the network is trained end-to-end, we freeze the weights and extract features from the last layer of the *backbone* i.e. 1280-dimensional features. The train and test splits remain the same throughout the pipeline. The features extracted by the fine-tuned network are now ready to be further optimized via feature selection, which we describe in the next section.

5.3.2 Information Gain based Feature Ranking

The features extracted from a deep neural network are of a high dimension (1280 in this case) and hence are often prone to redundancies and noise, which harms classification performance. To alleviate this, feature selection is often employed that aims at selecting the most important feature subset that can yield superior performance. This not only increases efficiency by reducing the feature space but also improves downstream performance.

In our proposed work we leverage Information Gain (IG) [135] as a quantitative metric for ranking the features. IG is a statistical method that computes the information gained for each feature in the dataset with respect to the target labels,

quantified in terms of the difference in the entropy [134] across targets with the conditional entropy upon the given feature. The effectiveness of using IG lies in its simplicity, inexpensive computational cost and fundamental statistical roots, which ensure its robustness.

The algorithm of IG, along with the relevant functions for entropy calculation have been formally described in Algorithm 2, Algorithm 3 and Algorithm 4.

Algorithm 2 $Information\_Gain\_Ranking(F, L)$

Input: Feature set F, Target labels L
Output: Feature scores F_S as list of tuples (f_i, s_i)
$F_S \leftarrow [\,]$ ▷ Initialize empty list
for each f_i in F **do**
  $H(L) \leftarrow Compute\_Entropy(L)$
  $H(L|f_i) \leftarrow Compute\_Conditional\_Entropy(L, f_i)$
  $IG(L; f_i) \leftarrow H(L) - H(L|f_i)$
  Append $(f_i, IG(L; f_i))$ to F_S
end for
$F_S \leftarrow ReverseSortedIG(F_S)$ ▷ Sort in descending order of IG
return F_S

Algorithm 3 $Compute\_Entropy(X)$

Input: Variable X
Output: Entropy $H(X)$
$H(X) \leftarrow 0$
for each unique x in X **do**
  $p(x) \leftarrow$ Probability of x in X
  $H(X) \leftarrow H(X) - p(x) \cdot \log(p(x))$
end for
return $H(X)$

Algorithm 4 *Compute_Conditional_Entropy*(X, Y)

Input: Variables X, Y (condition)
Output: Conditional Entropy $H(X|Y)$
$H(X|Y) \leftarrow 0$
for each unique y in Y **do**
  $p(y) \leftarrow$ Probability of y in Y
  $X_y \leftarrow$ Subset of X where $Y = y$
  $H(X_y) \leftarrow Compute\_Entropy(X_y)$
  $H(X|Y) \leftarrow H(X|Y) + p(y) \cdot H(X_y)$
end for
return $H(X|Y)$

Thus, from this step, we obtain a sorted list of features, sorted in descending order of their importance. We then extract the top-k% subset (k=50%, following [133]) as the final feature set which is to be used for the final classification. Note that train and test splits are maintained throughout to ensure there is no leakage of data.

5.3.3 Classification using Nearest Neighbours Classifier

Upon getting the final feature set, we employ a k-nearest neighbour (KNN) classifier on the features for final classification. The KNN algorithm is a simple non-parametric classifier that uses Euclidean distance metric to compute pairwise distances of the test samples with the training samples, and based on majority voting across 'k' closest samples determines the class of the test sample. Usually, the value of 'k' is a hyperparameter that needs to be tuned based on the dataset where it is being used. It is intuitive enough that 'k' determines the degree of confidence (i.e. how many reference points are chosen to infer a category), which, for a small dataset should be smaller as otherwise it would lead to class collisions. Similarly, 'k' should be increased for larger datasets. Following this, we chose k=5 for JAFFE and CK+ datasets (which are small) and k=100 for the larger FER2013 dataset.

5.4 Experiments

5.4.1 Datasets

The following standard FER datasets, all publicly available, have been used to evaluate the proposed framework suitably:

1. **JAFFE:** The Japanese Female Facial Expression, or JAFFE database [121] comprises 213 facial images spanning 7 emotion classes. All images here are of size 256x256 pixels.

2. **CK+:** The extended Cohn-Kanade dataset [122] consists of 981 images divided among 7 emotion classes, with each image being 48x48 pixels.

3. **FER2013:** This is the most challenging and one of the largest FER datasets till date [123], with 35,887 face image samples, categorized into 7 classes of emotions (shown in **Table 1**). All images are of 48x48 pixels. The reason for its difficulty lies in the presence of several labelling errors in its test set, affirmed by prior works [118]. Nevertheless, following relevant literature we evaluated our method on this dataset as well to investigate its generalizability.

5.4.2 Implementation Details

The proposed model has been implemented in PyTorch [151] using an instance of Kaggle [152], which provides a 16GB Nvidia Tesla P100-PCIE GPU. For the CNN-based transfer learning phase, all images were resized to 224x224 and standard data augmentations such as color jitter were applied to cater to the scanty data availability. The EfficientNet [20] CNN encoder was initialized with ImageNet [124] pre-trained weights, with a dense layer attached on top of it acting as the classifier head. The network was then fine-tuned end-to-end with the standard cross-entropy loss on the respective FER datasets, using ADAM optimizer [153] with a learning rate of $5e-4$ for 100 epochs. For JAFFE and CK+, the minibatch size considered was 4, while for FER2013 it was fixed to 32. For the feature selection phase, the IG-based algorithm described earlier is run on the deep features extracted by the pre-trained CNN to rank them, from which a top-k% subset is selected for the final classification. the KNN classifier [136] has been used for all classification evaluations performed in the proposed framework.

5.4.3 Evaluation Metrics

Four standard classification metrics, namely Accuracy, Precision, Recall and F1-score have been considered to evaluate the performance of the proposed model on the FER datasets. We have also provided confusion matrices obtained corresponding to the evaluation results to better interpret class-wise performance.

Table 5.2: Results obtained at each of the two steps of the proposed model, across each fold of the 5-fold cross-validation on the JAFFE dataset. All scores are expressed as percentages (%).

	Deep Features				After Feature Selection			
Fold	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
1	96.32	95.71	96.02	95.71	98.70	98.57	98.63	98.57
2	96.32	95.71	96.01	95.71	97.14	97.78	97.45	97.14
3	94.28	94.47	94.37	94.28	98.70	98.57	98.63	98.57
4	95.71	96.32	96.01	95.71	100	100	100	100
5	94.28	94.47	94.37	94.28	97.92	98.57	98.63	98.57
Average±Stddev	95.38±1.03	95.34±0.83	95.36±0.89	95.14±0.78	98.05±1.06	98.28±0.80	98.16±0.90	98.28±1.01

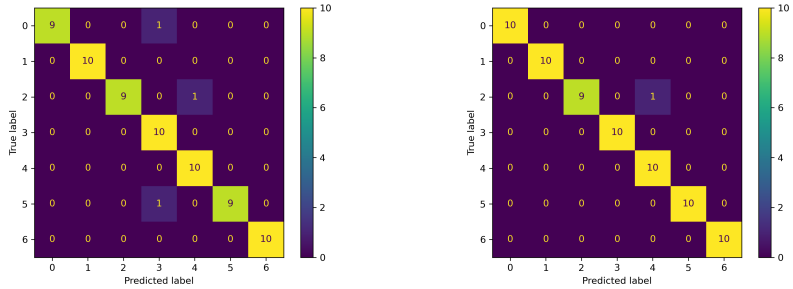
5.5 Results and Discussion

In this section, we provide and discuss quantitative evaluations conducted on three publicly available facial expression datasets which have been frequently used in prior literature [119], [132], [149]. For JAFFE and CK+, we employed a five-fold cross-validation scheme. For FER2013, since the train-test splits were already provided, we ran our method with five different random seed values.

5.5.1 JAFFE dataset

In Table 5.2 we show the results obtained on each fold of the cross-validation scheme on JAFFE dataset [121]. As clearly seen, the proposed model performs well across all folds, and the inclusion of feature selection consistently improves the scores by $\approx 3\%$ (also achieving 100% accuracy on one fold). We conjecture that reducing feature redundancies by feature selection is the cause of the observed performance gain. Furthermore, being a relatively small dataset, even single instances of erroneous predictions can lead to a high variation in the accuracy scores, which goes on to explain the high standard deviation scores. However, this does not undermine the promising empirical performance obtained by the proposed framework.

To analyse deeper into the classwise performance of the framework, confusion matrices have been depicted in Figure 5.3, which inform us how the model succeeds/fails in accurately predicting one class from another. We notice that class index '3' (i.e. "Neutral") is mistakenly predicted for true classes "Angry" and "Surprise" by the raw deep features, which after feature selection gets mitigated. This further proves the effectiveness of leveraging feature selection to improve the performance of the model. Finally, the emotion class-wise metric scores have been shown in Figure 5.4.



(a) EfficientNet features (b) After feature selection

Figure 5.3: Confusion matrices obtained upon classification on JAFFE dataset using (a) deep features extracted by EfficientNet and (b) mutual information-based ranking feature selection.

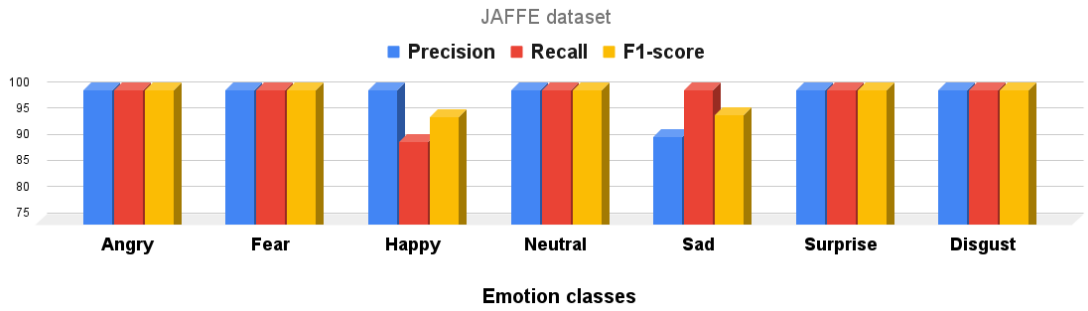
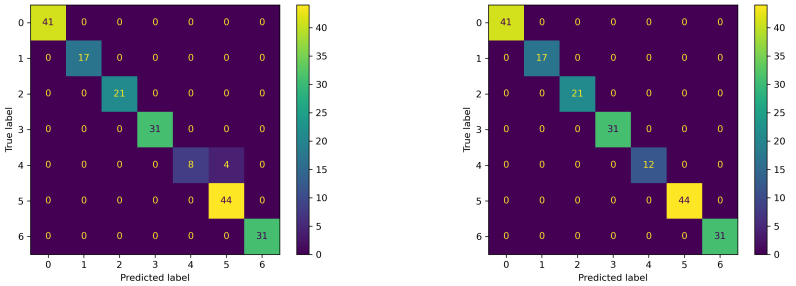


Figure 5.4: Emotion class-wise scores achieved by the proposed framework on JAFFE dataset.

Table 5.3: Results obtained at each of the two steps of the proposed model, across each fold of the 5-fold cross-validation on CK+ dataset. All scores are expressed as percentages (%).

	Deep Features				After Feature Selection			
Fold	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
1	98.81	95.23	96.51	97.97	100	100	100	100
2	98.24	95.07	96.14	97.46	100	100	100	100
3	96.28	97.51	96.96	96.95	99.14	98.81	98.98	99.49
4	98.81	95.24	96.52	97.97	100	100	100	100
5	99.47	98.81	99.07	98.47	100	100	100	100
Average±Stddev	98.32±1.22	96.37±1.70	97.04±1.17	97.76±0.58	99.83±0.38	99.76±0.53	99.80±0.46	99.89±0.22



(a) EfficientNet features (b) After feature selection

Figure 5.5: Confusion matrices obtained upon classification on CK+ dataset using (a) deep features extracted by EfficientNet and (b) mutual information-based ranking feature selection.

5.5.2 CK+ dataset

Table 5.3 outlines the results obtained by the proposed model under a five-fold cross-validation scheme on the CK+ dataset [122]. It is noteworthy to observe that the incorporation of feature selection boosts the performance to near perfection (99.89% accuracy on average of all folds). Additionally, the standard deviation among the folds is less compared to what we observed in JAFFE, showing more robust results with less variability obtained in this case.

From the confusion matrices in Figure 5.5 and the more concise class-wise metric scores shown in Figure 5.6, it is clear that the proposed model performs exceedingly well uniformly across all classes. Furthermore, the improvement due to feature selection can be clearly seen from the two confusion matrices, where the second step of the pipeline seems to eradicate confusion between classes “Sad” and “Surprise”.

5.5.3 FER2013 dataset

In addition to the above FER benchmarks, we also evaluate our method on the very popular FER2013 database [123], which is to date one of the largest

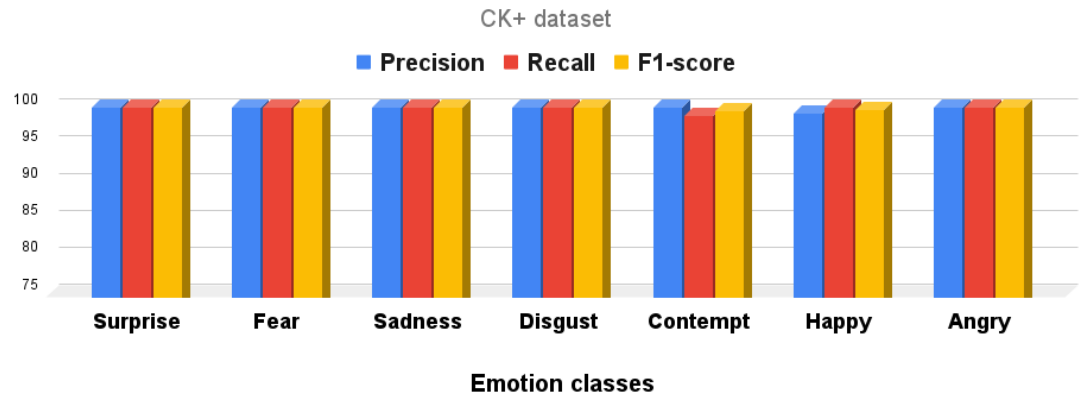


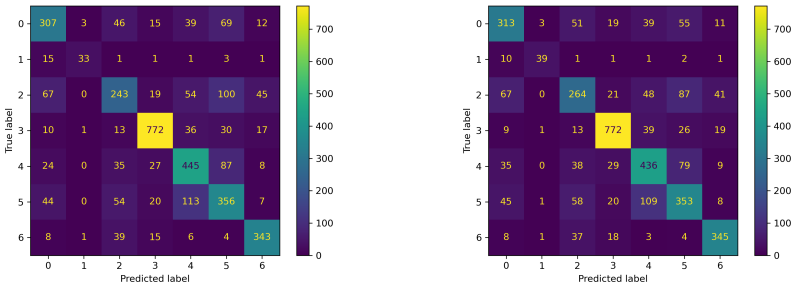
Figure 5.6: Emotion class-wise scores achieved by the proposed framework on CK+ dataset.

labeled facial expression corpus. In addition to its volume, this dataset also has a lot of noisy or incorrectly labelled images (as highlighted in prior works [119]), which makes it a very challenging benchmark even after a decade since it was first released.

The FER2013 dataset comes with demarcated train-validation-test splits, so a cross-validation scheme is not applicable. Instead, we resort to running our entire pipeline under 5 different random seed initializations and report the findings. We have tabulated these in Table 5.4. Here, we see that although feature selection boosts performance compared to raw deep features, the empirical gain ($\approx 1.5\%$) is not as much as what we had obtained in JAFFE or CK+. This maybe due to the inherent complexity of the dataset. However, once again we achieve consistent trends as in previous results, asserting the usefulness of the proposed method. For inter-class performance analysis, we have provided confusion matrices in Figure 5.7 and a bar chart depicting scores across different emotion classes in Figure 5.8. It may be inferred that the model has a lot of scope to improve on FER2013 empirically; however, when we contextualize the performance with prior state-of-the-art works (comparison in Sec. 5.5), it becomes evident that our model, with its simplicity and lightweight characteristics, is highly competitive to more complex and computationally heavy approaches.

Table 5.4: Results obtained at each of the two steps of the proposed model, across each of 5 randomized seed runs on FER2013 dataset. All scores are expressed as percentages (%).

	Deep Features				After Feature Selection			
Runs	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
1	70.11	68.91	69.46	70.05	71.25	70.47	70.94	71.38
2	69.88	67.80	68.85	68.71	71.16	69.21	70.02	70.27
3	70.14	68.67	69.11	69.63	71.76	70.12	70.78	70.77
4	69.75	67.52	68.66	68.24	70.07	68.85	69.36	70.14
5	69.71	68.37	68.97	69.82	70.84	69.21	70.02	71.08
Average±Stddev	69.92±0.20	68.26±0.59	69.01±0.30	69.29±0.78	71.02±0.62	69.57±0.69	70.22±0.64	70.73±0.53



(a) EfficientNet features (b) After feature selection

Figure 5.7: Confusion matrices obtained upon classification on FER2013 dataset using (a) deep features extracted by EfficientNet and (b) mutual information-based ranking feature selection.

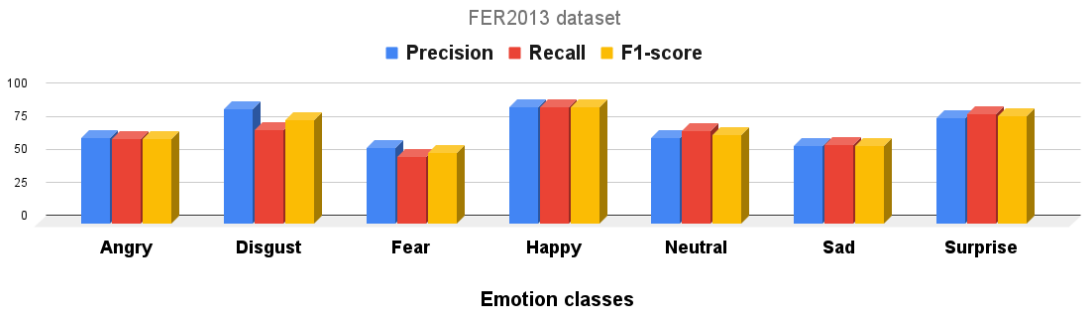


Figure 5.8: Emotion class-wise scores achieved by the proposed framework on the FER2013 dataset.

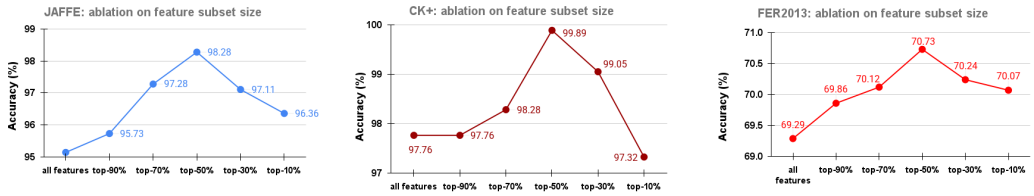


Figure 5.9: Ablation study on feature subset size selection.

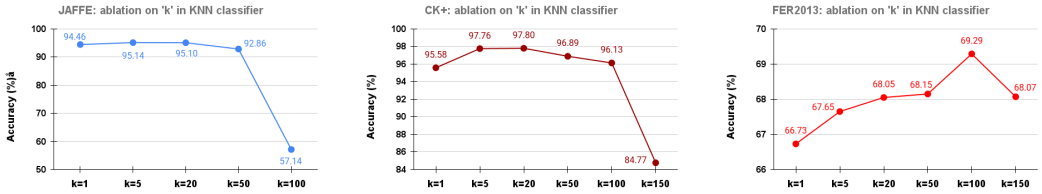


Figure 5.10: Ablation study on choice of 'k' for KNN classifier. The results are reported on the extracted raw deep features.

5.5.4 Ablation Study

In this subsection, we perform an ablation study on the choice of selection of the size of the subset of features (set to the and the 50% for the main results) and the value of 'k' in the classifier of the k-nearest neighbor in the original deep features.

5.5.4.1 Feature subset size selection

Based on the feature rankings obtained via the IG-based feature selection algorithm, we choose various sizes of the top-ranked features, i.e. going from from top-90% to top-10% of the features, for each of the facial emotion recognition datasets. We show the accuracies obtained graphically in Figure 5.9.

From Figure 5.9, we observe that while removing unimportant features boost performance (which is precisely the objective of feature selection), reducing the feature space too liberally hurts the model's performance. While an exhaustive search over the entire range of features is infeasible, in particular we observe a peak around the 50% mark. This justifies our choice of using the top-50% feature subset for the final classification. Notably, a prior work [133] also found the same feature subset size to be optimal, which also supports our choice.

5.5.4.2 Value of 'k' in KNN classifier

We also ablate the number of neighbors chosen for classification in the KNN classifier, on the raw deep features extracted from the transfer learning stage, for each dataset used in this study. The results are depicted in Figure 5.10.

Table 5.5: Comparison of the proposed approach with some of the prior state-of-the-art works in literature across JAFFE, CK+ and FER2013 datasets. Note that only accuracy values have been compared as it is the most obvious and only metric reported in all of the works.

JAFFE		CK+		FER2013	
Method	**Accuracy** (%)	**Method**	**Accuracy** (%)	**Method**	**Accuracy** (%)
Ghosh et al. [132]	89.61	Zhang et al. [154]	97.20	Tang et al. [155]	64.41
Zhang et al. [154]	90.95	Minaee et al. [92]	98.00	Minaee et al. [92]	70.02
Minaee et al. [92]	92.80	Kumari et al. [118]	98.01	Kalsum et al. [156]	69.87
Khattak et al. [119]	95.65	Foggia et al. [157]	92.52	Fard et al. [158]	68.25
Sultana et al. [159]	93.75	Sultana et al. [159]	94.84	Kumar et al. [160]	71.16
				Kong et al.	70.63
Ours	**98.57±1.01**	*Ours*	**99.90±0.23**	*Ours*	**70.73±.0.53**

From the plots in Figure 5.10, it is evident that a lower value of 'k' is desirable for JAFFE and CK+, owing to their smaller dataset size. For JAFFE, k = 5 gives the best performance. Although k = 20 is the most optimal for CK+, we point out that its gap with k=5 is trivial (= 0.04). Increasing 'k' will increase run time of the model, so choosing a lower value of 'k' is desirable, justifying our usage of k=5 for both these datasets. For FER2013, owing to its considerably larger size, a greater value of 'k' leads to best performance, i.e. k=100 (where it peaks). This ablation study justifies the choice of 'k' in all experiments.

5.5.5 Comparison to state-of-the-art

We have also compared the empirical performance of our proposed framework against several prior works in literature on each of the datasets used. The results are shown in Table 5.5. On JAFFE [121] and CK+ [122], we outperform all prior arts by significant margins, including almost near-perfect performance on the latter dataset. It is worth noting that few of these works employ high-computational and complex approaches such as attention mechanism in Minae et al. [92] and doubly channelled LSTM networks [154]. In contrast, our approach is simpler, lightweight and still robustly outperforms these complicated methods. For FER2013 [123], the proposed method is highly competitive to prior state-of-the-art approaches, outperforming several of them. It is well known that FER2013 is a lot more challenging and noisy dataset and hence, achieving competitive performance on it is reasonable evidence of the robustness our method provides. To sum up, our method achieves sound and consistently robust results across both small-scale and large-scale facial expression datasets, highlighting its effectiveness and generalizability.

5.6 Summary

In this work, we explored a lightweight, computationally inexpensive, yet robust and strong pipeline for facial expression recognition, harnessing the power of transfer learning and statistical feature selection strategies. Upon thorough evaluation on three publicly available FER datasets, we observed that the model achieved strong and competitive results, outperforming several existing state-of-the-art works in literature. The model being computationally efficient can be deployed for real-time and edge device applications. Furthermore, our pipeline is very generic and can adopted for different tasks in computer vision. We wish to expand our study and investigate feature redundancies in more detail for facial images to understand which features yield redundancies and from which regions of the face do they come from, which we believe could lead to development of stronger FER models in future.

CHAPTER 6

Emotion Landscapes in Group using Image Processing and Shallow CNN

Emotion Recognition is a key aspect of any Human-Machine Interaction (HMI) system. Recognizing emotions correctly lets HMI systems determine the right replies in order, based on the context and the emotion shown by the person or people. Deep Learning with Deep Neural Networks (DNN) has achieved amazing progress in picture classification and face detection, even surpassing human accuracy. Numerous publications have documented the effective utilization of Deep Neural Networks, such as Convolutional Neural Networks. CNN is now the most common technique for classifying facial images since it combines the procedures of feature extraction and classification into one mathematical model. They learn the features that are wanted on their own from the input photos and have been shown to be strong against changes in face image data. But there is one huge problem with CNNs: the models that are particularly accurate have a lot of hidden layers, which means they are very deep and need a lot of computational power, memory space, and time to train themselves.

Our research proposes two experimental methodologies that can significantly enhance the fields of CNNs and HMI systems. We were able to get extremely good results with the first method, which used a shallow CNN with only three and four hidden layers to classify emotions. We were able to do this because we meticulously processed the input images through a series of Image Preprocessing algorithms before giving them to the CNN for training. For the second method, we created the interpretation of Emotion Landscape, which shows how emotion classes are spread out in static images or videos with a lot of people's faces in them. This resembles the Group-Level Emotion Classification research

and literature, although it differs in terms of potential applications. The reason for this interpretation is to promote the idea of looking at how people in a group setting express their emotions and how they affect each other, to see how the distribution of emotions changes over time and to create an emotional landscape over time, and to better understand how people express their collective feelings non-verbally through facial expressions in social situations.

6.1 Introduction

Emotions are an important and necessary part of being human. Emotions can serve as a powerful means of communication, conveyed through facial muscles, hand movements, body gestures, and vocal modulations. This paper concentrates on the visual perception of emotions conveyed through an individual's facial expressions. According to Mehrabian (1971), facial expressions account for 55

The aim of the proposed work is to explore a novel viewpoint on the problem domain of Facial Emotion Classification and examine it within the context of social groupings. The current literature regarding the applications of Machine Learning and Deep Learning predominantly emphasizes the training and evaluation of models using labeled facial photos. We aimed to investigate the concept of recognizing emotions shown by individuals in the presence of others, specifically within a group-level context. Section 3.6 talks about our idea for an emotional landscape for a group of people in a social setting. This is partially based on the examination of emotion inference from a group-level image. Barsade et al. (1998) and Kelly et al. (2001) discuss the top-down and bottom-up components of emotion inference. The top-down component aims at the collective emotion of a group, derived from the distinctiveness of each member's facial expressions, whereas the bottom-up component initiates at the group level and subsequently addresses the emotions of individual members. For a static group-level image, we get the emotion distribution for each of the front-faces that we found. The EmotiW challenge has three classes: "positive," "neutral," and "negative." We think that this could limit the kinds of emotional information that CNNs can gather, thus we used seven emotion classes for our study: surprise, sadness, fear, distress, anger, disgust, and happiness. You may find the distribution of values for these seven classes for the group-level image by using an aggregation function on the data for each of the detected faces. When we look at a video of the same group, we can get a time-series distribution of the emotion classes values. This may be shown as a landscape that changes over time because of different people's emotions and the emotions that the group thinks are happening.

6.2 Methodology Adopted

6.2.1 Problem Statements

The initial section of this work delineates a cost-effective method to attain satisfactory accuracy rates utilizing CNNs with only 3 and 4 hidden layers. We can think of the CNN architecture like this:

$$x^1 \rightarrow\ w^1 \rightarrow\ x^2 \rightarrow \ldots \rightarrow x^{n-1} \rightarrow\ w^{n-1} \rightarrow\ x^n \rightarrow\ w^n \rightarrow z \tag{6.1}$$

The CNN gets its input from x^1, which is a 3D tensor, or image. A tensor w^i represents all the parameters that each layer uses to process data. In this case, the CNN will be used to sort images into 7 classes or values in z. The x^n is changed into a C-dimensional vector, and the i-th element of that vector holds the posterior prediction $P(C_i|x^1)$. We want you to know that in this study, we treat the convolution operation, pooling operation, and normalizing operation as one layer in the CNN. There are also Dense Layers and the last output layer in the CNN. The CNN does not treat the drop out layer as a separate layer.

The second portion of this study is about making sense of the CNN's results in a way that makes sense. We discover z^i for each of the faces we see in the group-level image of people's faces. In this case, z will be a column vector with 7 elements. These z numbers can be thought of as the chances that each of the 7 emotion groups is true. Adding up the probability values for all the identified faces in each of the 7 emotion classes will give you z', a column vector that gives you a global context-free probability value for each class. If a single group-level image shows m faces, then:

$$f\left(\begin{bmatrix} z_0^0 \\ z_1^0 \\ \vdots \\ z_5^0 \\ z_6^0 \end{bmatrix}, \begin{bmatrix} z_0^1 \\ z_1^1 \\ \vdots \\ z_5^1 \\ z_6^1 \end{bmatrix}, \ldots, \begin{bmatrix} z_0^{m-1} \\ z_1^{m-1} \\ \vdots \\ z_5^{m-1} \\ z_6^{m-1} \end{bmatrix}\right) = \begin{bmatrix} z'_0 \\ z'_1 \\ \vdots \\ z'_5 \\ z'_6 \end{bmatrix} \tag{6.2}$$

Here f is an aggregation function such that $z'_i = \frac{\lceil \sum_j^{m-1} z_i^j \rceil}{\sum_i^6 \lceil \sum_j^{m-1} z_i^j \rceil}$

6.2.2 Workflow

We got the JAFFE and Cohn-Kanade facial picture databases from their official sites. You can download them for free after saying how you want to use them and getting permission from the database authors. The CNN was trained using the

individual facial images from these two datasets. During training and testing, every facial image went through a customized pipeline that extracted certain or targeted areas of the face. We then min-max normalized these ROIs to put the pixel values between 0 and 1. This helps the CNN learn quicker by making it converge faster. The CNN's output column vector z was set into a range of 0 to 100. So, each of the 7 values in z could be seen as a probability value for an emotion class. A grey-scaled group-level image with many visible faces is sent to the pre-processing pipeline that was talked about earlier during the testing phase. There were target ROIs for each face i that was found in the image. The trained CNN was given these ROIs one at a time, and it gave back a z vector for m recognized faces. The m output vectors were then given to the aggregation function f to get z', which shows the global probability values for each of the 7 emotion classes in the group-level image.

6.3 Implementation

6.3.1 Datasets

We used the JAFFE facial emotion image dataset from [161] in the beginning of our experiments. The Japanese Female Facial Expression (JAFFE) database has 213 pictures of 7 different facial expressions, including 6 fundamental ones and 1 neutral one. These images were taken of 10 Japanese women. 60 Japanese people assessed each picture on 6 different emotion classes. We utilized the dataset to make our first CNN model and test the pre-processing pipeline.

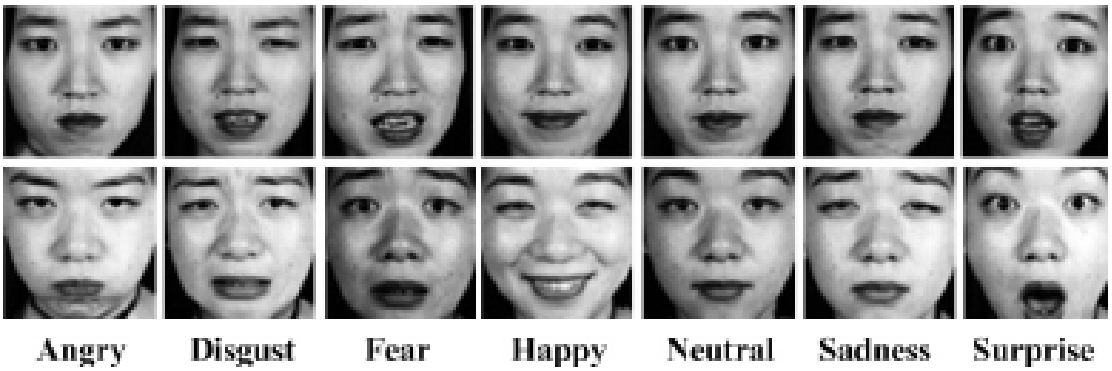


Figure 6.1: Two faces from the JAFFE dataset and the six ways they show emotion.[74][121][161]

The Cohn-Kanade AU-Coded Facial Expression Database [162] is a standard dataset for research in automatic facial image analysis and emotion prediction. It has more pictures than the JAFFE dataset, therefore it was employed in the later experimental stages to make changes and improvements to the pipeline and CNN

models. There are two versions of the Cohn-Kanade dataset. This is version 1. There are 486 sequences from 97 posers in it. There is a neutral expression at the start of each sequence, and it builds to a peak expression.

Figure 6.2: Examples of facial expressions from the Cohn-Kanade dataset.[79][122]

6.3.2 Workstation Configuration

The workstation that was utilized to pre-process the image dataset had an i7-4790K quad-core processor with 8 threads, a maximum turbo frequency of 4.4GHz, and an 8MB smart cache. The graphics processing unit is the NVIDIA GeForce GTX 1080, which has 8GB of GDDR5 memory, a boost clock speed of 1733MHz, and 2560 CUDA cores. The computer had 32GB of DDR4 RAM. We chose Linux Mint 18.3 with the Cinnamon desktop since it was fast and easy to use. The training and testing on the GPU indicated above were done reliably with CUDA 9.0.

We utilized Python 3.5.1 and the Keras Library for Python to make our prototypes. Keras is a high-level deep learning API that makes it easier to build conventional neural network architectures like CNNs and RNNs. It also has a lot of optimizers, classifiers for neural network output layers, weight matrix initializers, and more to pick from.

OpenCV [163] API for Python is an open source Computer Vision framework that has all the tools and algorithms you need to work with image data. OpenCV, along with the C++ multi-purpose package dlib and imutils (to support dlib), enabled us find and clip out faces and get the ROIs we need as input features for the CNN model.

6.3.3 Detecting Faces in Images

OpenCV has two pre-trained classifiers for detecting faces: the Haar-Cascades [164] and the Local Binary Pattern (LBP)-Cascades [165]. OpenCV comes with these classifiers already trained on a lot of positive (pictures with faces) and negative (images without faces) images. Table 6.1 compares the two classifiers and shows how they work.

Algorithm	**Advantages**	**Disadvantages**
Haar	1. High detection accuracy 2. Low False Positive Rate	1. Computationally complex and slow 2. Longer Training Time 3. Less accurate on black faces 4. Limitations in difficult lighting conditions 5. Less robust to occlusion
LBP	1. Computationally simple and fast 2. Shorter training time 3. Robust to local illumination changes 4. Robust to occlusion	1. Less accurate 2. High false positive rate

Table 6.1: Comparison of Haar and LBP classifier algorithms

The main distinctions between them seem to be how fast and how accurate they are. The LBP Cascade Classifier is quicker, but the Haar-Cascades is more accurate. We utilized the Haar-Cascades to find faces in group-level photos because we care more about accuracy than speed. When you run a group-level snapshot through the Haar-Cascade Classifier, you get the result in Figure 3:

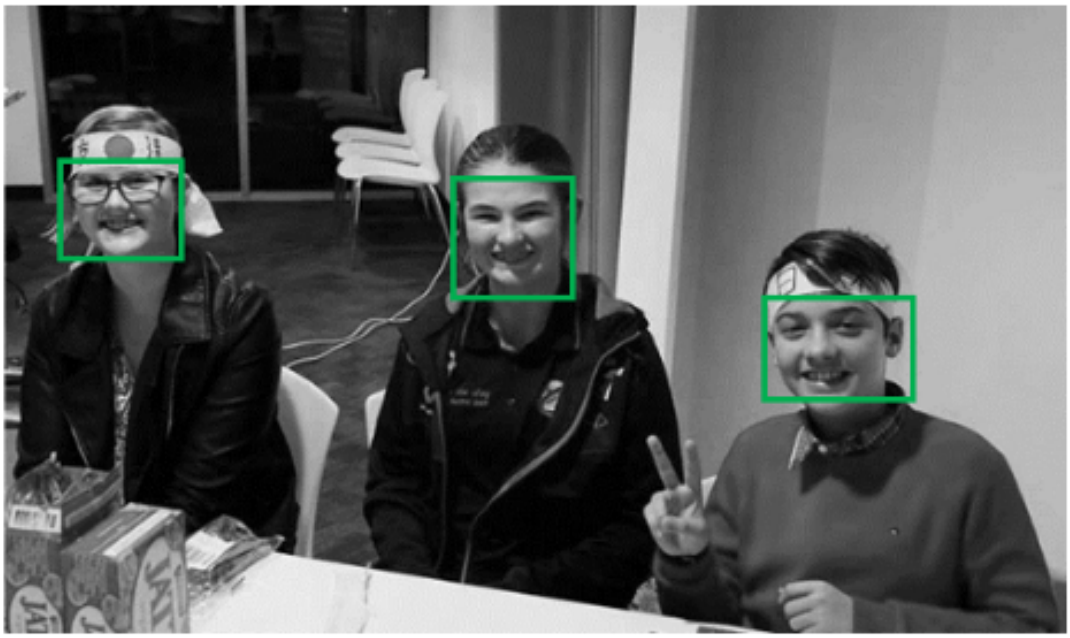
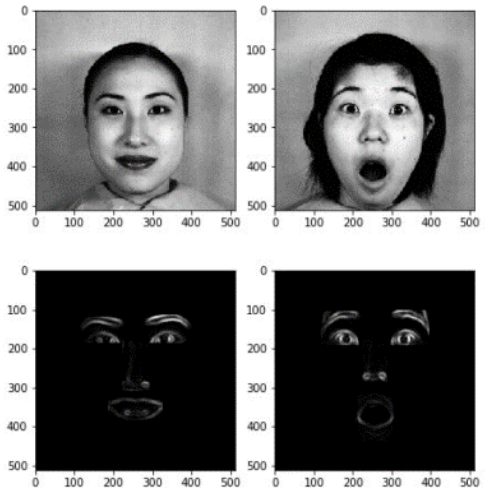


Figure 6.3: A sample image from our experimental test dataset with three faces found and indicated in green boxes.

6.3.4 Targetted ROI for feature extraction

We utilize the dlib facial feature landmarks to find the areas around the eyes, lips, and nose from the group-level images of the faces we found. We take these exact ROIs because we think that these traits are enough to tell what emotion a person's face is showing. Dlib noted the face markers in Figure 6.4.a below. Figure 6.4.b shows the referential landmark coordinates that the dlib face landmark predictor gave. It was trained using the 68-point iBUG 300-W dataset [166]. There are alternative detectors for face landmarks, so keep that in mind. You can get a more thorough 194-point model that was trained on the HELEN [167] dataset. The dlib library enables us ignore the details and focus on finding and cutting out the ROIs we want, no matter which dataset or landmark predictor we employ.



(a) ROIs as marked and extracted by dlib.

(b) Visualizing the 68 facial landmark coordinates from the iBUG 300-W dataset.

Figure 6.4

6.3.5 Shallow CNN to identify emotions expressed by detected faces

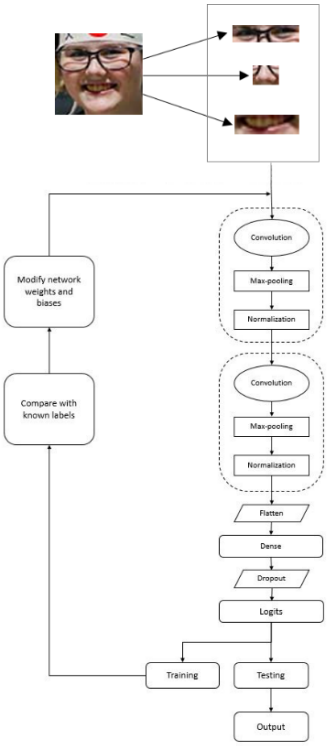


Figure 6.5: Shallow CNN architecture we used.

For our experiments, we employed a CNN model with three hidden layers. The model went through 50 epochs of training. Figure 6.5 demonstrates how the CNN models take the target ROIs that were taken from each recognized face as

input. Our CNN models have a straightforward structure. They have two repeated sequences of convolution, max-pooling, and normalization operations, followed by one or two densely linked neuron layers, and finally, a classifier in the last output layer. We decided to use such a minimal CNN architecture since we think we have already narrowed down the image dimensions to the areas where we want the CNN to learn. You might also say that this makes it easier for the network to learn how to map the input to the output classes by lowering the amount of parameters and adjustments it has to make. Adding more layers means adding more factors that need to be learned. This is useless and a waste of time because the input feature space is smaller.

6.4 Results

Figure 6.6: Test Image 01 [161]

Face	Dist.	Sad	Surp.	Happy	Fear	Neut.	Angry
1	0.0	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	1.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	1.0	0.0	0.0	0.0
5	0.0	0.0	1.0	0.0	0.0	0.0	0.0
6	0.0	0.0	1.0	0.0	0.0	0.0	0.0
Landscape	0.0	0.0	0.67	0.33	0.0	0.0	0.0

Faces Detected: 6
Counter: {Surprised: 4, Happy: 2}

Figure 6.7: Test Image 02

Face	Dist.	Sad	Surp.	Happy	Fear	Neut.	Angry
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Landscape	0.0	0.0	0.0	0.67	0.33	0.0	0.0

Faces Detected: 3
Counter: {Happy: 2, Fear: 1}

There is a correct interpretation of the CNN results on the right side of each figure. After 50 epochs of training, the CNN model had a validation accuracy of 94.28

Face	Dist.	Sad	Surp.	Happy	Fear	Neut.	Angry
1	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Landscape	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Faces Detected: 1
Counter: {Fear: 1}

Figure 6.8: Test Image 03

6.5 Emotional Landscape Evolution for Group-Level

We already talked about how to figure out the probability of different emotions in a picture of a group of people. The method can be directly applied to recordings of social events by passing each frame to the pipeline resulting in an output vector for the seven emotion classes for every detected face throughout the video. The output is really a 4D tensor with the following axes: face ID, time, emotion class label, and emotion class label probability. This tensor is not a useful approach to get further information or insights into the main group-level emotions of the people in the video. We suggest a more visual way to show the output tensor: an emotional environment that changes over time in the video. There are 7 emotion classes and time on the 3D surface plot, and each point on the surface is an accumulated function of them. We think that this idea is still in its early stages, but we believe that a landscape function surface like this could be very useful for making user-friendly HMI by adding the ability to understand and predict how users will feel after a long conversation. The HMI would be able to guess how most people would feel about a certain topic or answer, and it might choose to learn how to take the conversation along.

6.6 Summary

Our research proposes two primary methodologies for emotion categorization through facial expressions, facilitating a more human-like inference of others' expressions and emotional states in Human-Machine Interface (HMI) systems. The first concept is that if the images that are input are pre-processed in a way that renders sense, this lets the CNN that learns to do the classification task with those images have fewer or shallower hidden layers. Pre-processing the input image and removing context-specific portions to an assumption reduces the "space" that the

CNN must explore to map the input to the output classes. The second notion we talked about is to show how group-level emotions change over time as a landscape. This enables diverse interpretations of discrete classification labels that evolve over time and expands the opportunity to investigate and bridge the gap between human and machine interpretation. The primary constraint identified regarding our suggested concept is the impact of the time intervals at which the landscape is refreshed. Too short of an interval can give you wrong results since there is a limit to how quickly human facial expressions can change. Even yet, HMI systems that can watch how a person's facial expressions change in a social context might considerably improve their ability to make decisions and communicate to people.

CHAPTER 7

Constraints of FER : Mutual exclusivity of class

Emotion Recognition is an essential component of any Human-Machine Interaction (HMI) system. HMI systems can choose appropriate responses one after the other if they can correctly identify emotions based on the situation and the emotion shown by the person or people. Deep Learning with Deep Neural Networks (DNN) has achieved amazing progress in image classification and face detection, even surpassing human accuracy. But because of the stringent constraints on what classification means, these kinds of models can't be used very much in HMI systems. In the last ten years, Convolutional Neural Networks (CNN) have become the most popular type of deep learning for image-based classification problems. This is because they combine the phases of feature extraction and classification into one mathematical model. But we still don't have a CNN-based framework that is flexible enough to make an adaptive HMI system that is easy to use. In our research endeavors, we have experimented with pre-trained CNN for FER and endeavored to adapt it for video data, facilitating HMI systems to get a continuous stream of user emotions. We think this is the best way to make HMI systems that can respond in an intelligent way, although the method has difficulties like emotion classes being mutually exclusive and prediction noise. This research elucidates the challenges that HMI systems have while employing CNN for FER-related tasks utilizing video data . We have included one of our relevant experiments to show where and how they appear. In this experiment, we utilize two pretrained CNN models to do FER on a video of a group of individuals conversing. We also talked about some ways to solve the problems and get around them while still utilizing the same CNN models as before.

7.1 Introduction

Emotions are an important and necessary part of being human. Emotions can serve as a powerful means of communication, conveyed through facial muscles, hand movements, body gestures, and vocal modulations. This work concentrates on the visual perception of emotions conveyed through an individual's facial expressions.

This work presents a nuanced critique of the research and development of DNNs. DNNs as a Universal Function Approximator [168] have excelled in limited problem domains, although there is significant potential for expansion beyond their current training parameters. The EmotiW challenge has three classes: "positive," "neutral," and "negative." We think this might be too limiting for us to show off what we've seen. In Section 2, we discuss one of our own experiments in which we used pretrained CNNs with seven emotion classes: "surprise," "sad," "fear," "distress," "angry," "disgust," and "happy." This effectively shows the limitations of FER that have been noted in both academic literature and industrial applications. In Section 3, we talk about the constraints we saw in our experimental setting such that they are easy to understand in a practical approach. We end our research with suggestions for different methods to get around the problems we talked about.

7.2 Experimentation

In this section, we discuss about one of our tests that will later show how DNNs for FER don't work well in HMI systems. We will talk about our Neural Network Pipeline, which finds human faces in a movie, puts them in order, and tags them along the way. It then does FER on all of them.

7.2.1 Pipeline

The pipeline takes a video in a typical MPEG-supported format and turns it into a series of frame images, as shown in Fig. 7.1. Dlib's The Face Detection Model API from [169] lets us send each frame to the pretrained CNN, which finds and cuts out the faces of people in the frame. It is also possible for the API to identify faces based on how similar they are. We utilized this approach with a dictionary database so that every face the CNN found was compared to every other face in the database. If they were more than 75% identical, they were given a unique identifying number and the same selection of faces from the dictionary. After reading all the frames from the video, the pre-trained CNN from Octivaio et

al. [170] was given a dictionary of cataloged crops of human facial images. It then returned a column vector of softmax probabilities for 7 emotion classes: Happy, Sad, Neutral, Angry, Fear, Disgust, and Surprise. For each participant in the movie, the emotion categorization vectors for all of their faces were put together to make a time series matrix. The time series matrix is shown in Fig. 7.2. Each column was a softmax prediction vector for an emotion class for a person with ID i from a frame at timestamp t_i when that individual's face was found and recognized.

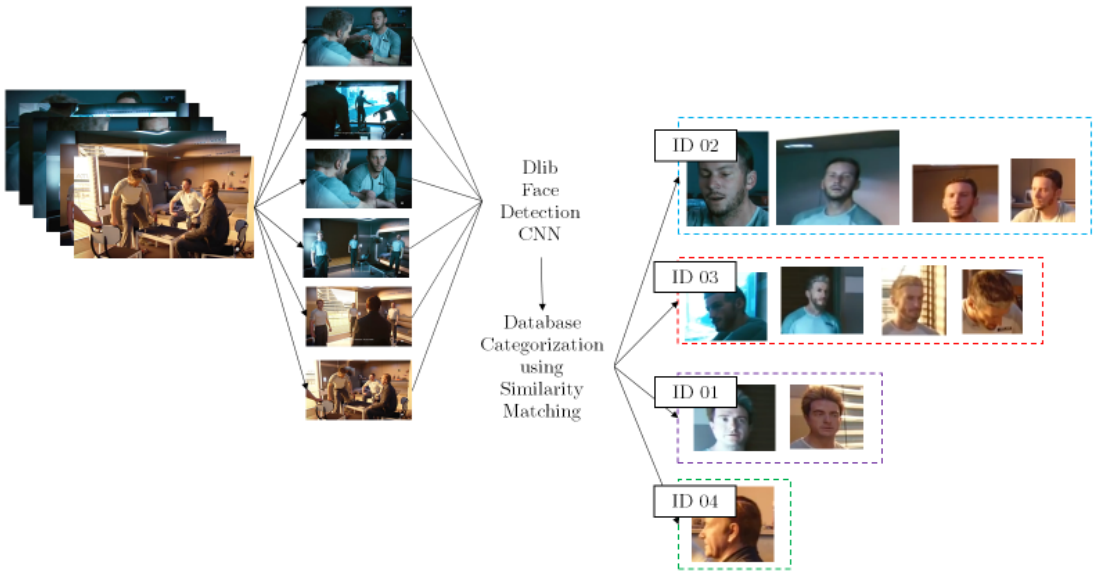
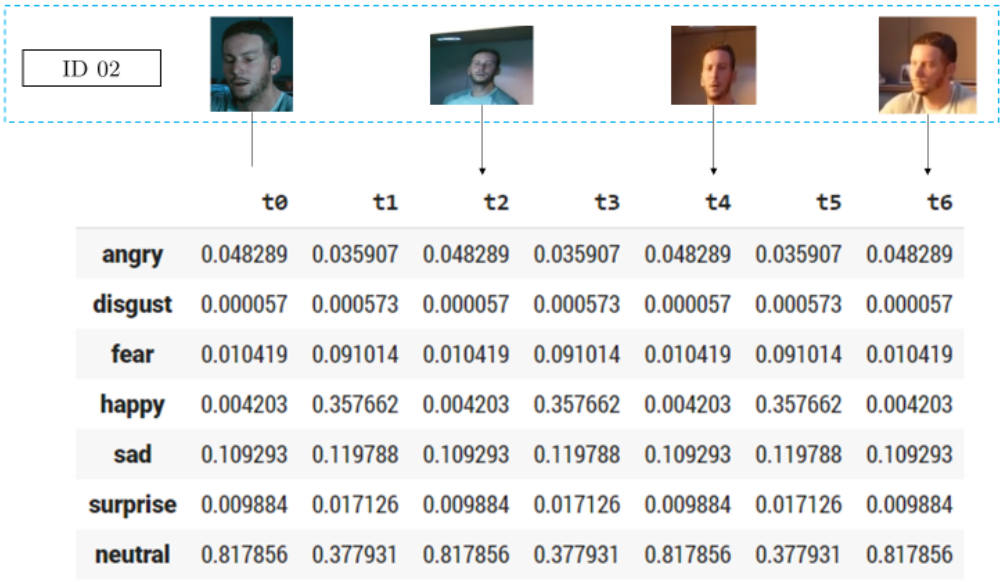


Figure 7.1: Experimental Pipeline used for video-based FER



	t0	t1	t2	t3	t4	t5	t6
angry	0.048289	0.035907	0.048289	0.035907	0.048289	0.035907	0.048289
disgust	0.000057	0.000573	0.000057	0.000573	0.000057	0.000573	0.000057
fear	0.010419	0.091014	0.010419	0.091014	0.010419	0.091014	0.010419
happy	0.004203	0.357662	0.004203	0.357662	0.004203	0.357662	0.004203
sad	0.109293	0.119788	0.109293	0.119788	0.109293	0.119788	0.109293
surprise	0.009884	0.017126	0.009884	0.017126	0.009884	0.017126	0.009884
neutral	0.817856	0.377931	0.817856	0.377931	0.817856	0.377931	0.817856

Figure 7.2: The concept of a time series matrix of emotion class predictions for a person with ID 01

7.2.2 Face Detection

Dlib [169] has released an open-source DNN model and Python API on GitHub that can find and identify faces. We can put some pictures of people's faces into this model so that we can find and identify them in real time. Then, we can use the pretrained CNN to find them in a new picture that we haven't seen before.

7.2.3 Emotion Classification

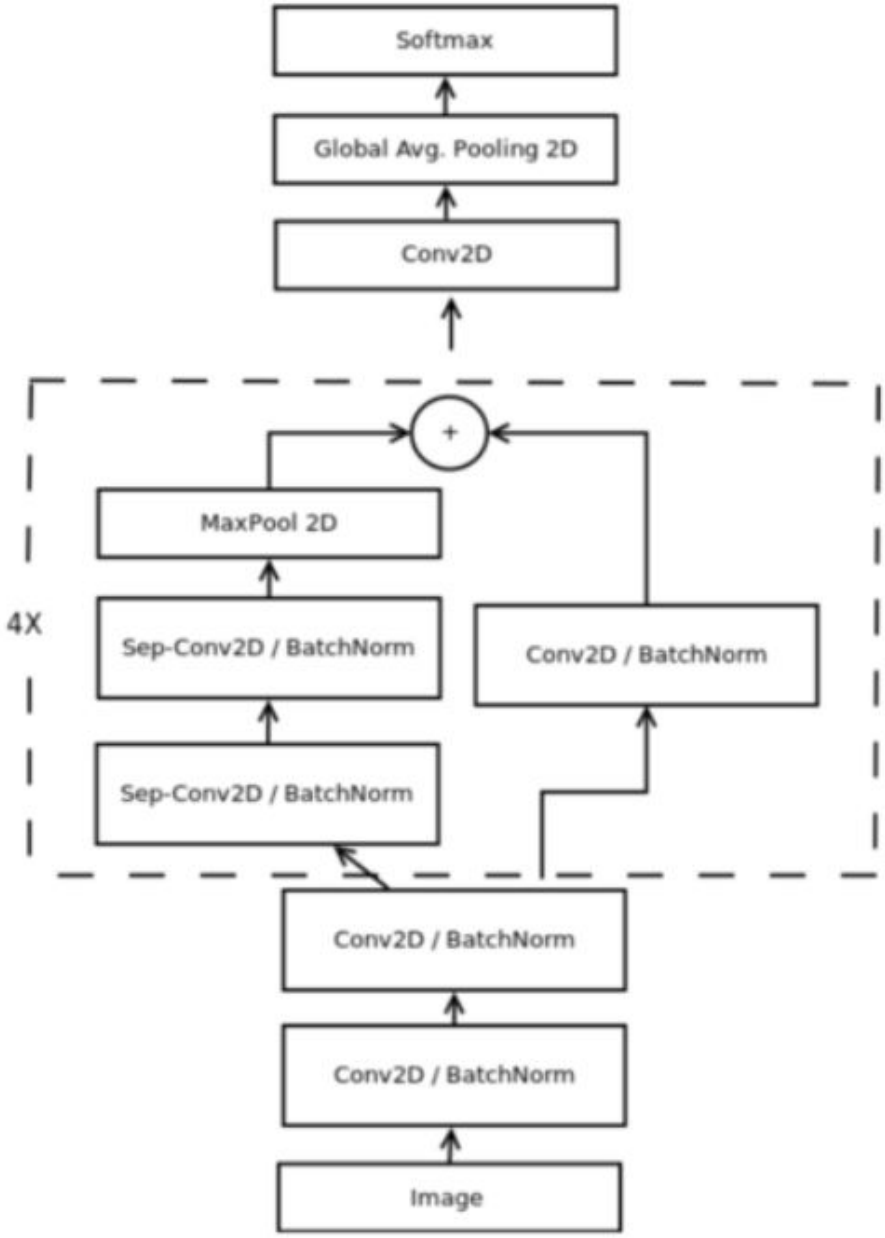


Figure 7.3: CNN architecture for Octivaio et. al. [170]

The authors, Octivaio et al. [170], have made the FER model available on Github under an open-source license. They have also published a paper about the same model that talks about their trials with FER and Gender Recognition tasks. The model has been trained using the FER 2013 [171] dataset for FER and the IMDB Gender Dataset [172]. The architecture is new and different because there isn't a Fully Connected Layer (FCL) anywhere before the Softmax Classification Layer at the end. The authors explain their choice to get rid of the FCL by saying

that they saw that 90% of the learnable parameters (neuron weights and biases) in the VGG16 [15], Inception V3 [173], and Xception [108] architectures are in the FCL layers. They suggested the architecture depicted in Fig. 7.1 to make the model smaller, speed up training, and keep a good degree of classification accuracy. This met all of our needs for our experiments because we intended to employ light-weight versions of CNN that can be natively deployed to run on and be incorporated into HMI devices.

7.2.4 Experimental Results and Discussions

We plotted the time series matrix introduced in Fig. 7.2 in 7 area plots for each of the 7 emotion classes. Each emotion plot has 4 overlapping areas for the 4 persons identified in the video, showing the probability of emotion experienced along the duration of the events as they unfold in the video.

We need to first grasp the clip scene in order to establish the stage and figure out what the plots are trying to say. The individual with ID 04 runs a private military firm that employs a specialist assault team of three troops, known by IDs 01, 02, and 03. Recently, they were sent to attack a terrorist stronghold and free a high-profile government figure who was being held for ransom. After a successful rescue mission, they rest in a room at an army camp and talk about the mission and what it means for them, as they think they might be called back to do another mission. The owner of the company, ID 04, comes in as they are talking and gives them a shot of whiskey, praising and thanking them for what they did.

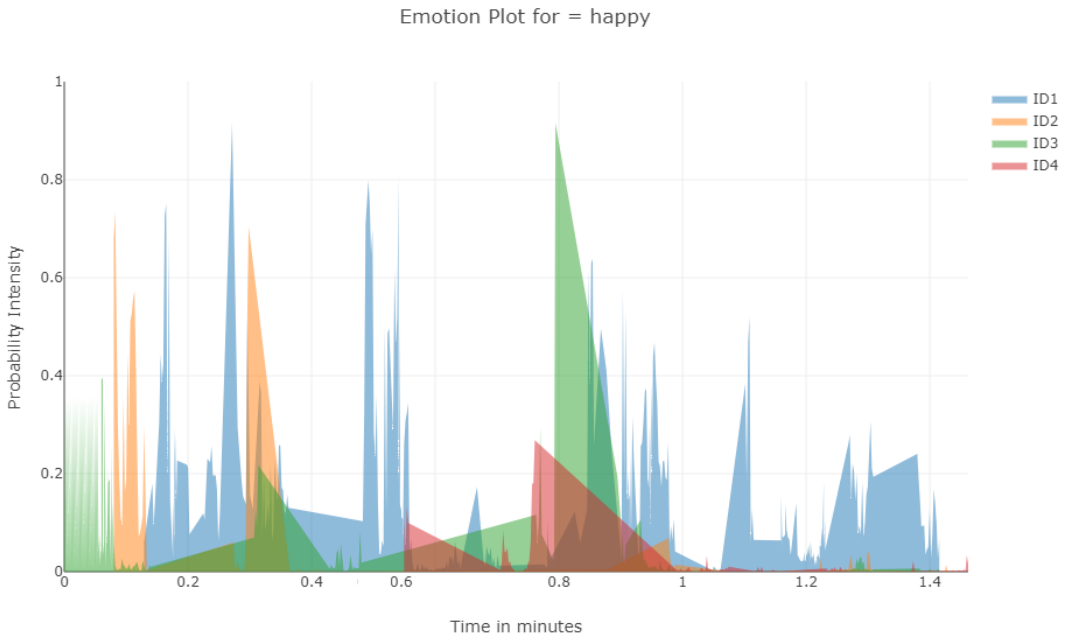


Figure 7.4: Area plot for emotion "Happy"

ID 01 is quite happy the whole time and has a big smile on their face. ID 02 looks more serious and thinks about how the mission could go wrong a lot and how the terrorists will respond when they attack their residence. ID 03's facial expressions are more subtle and professional, and he likes to make fun of ID 01 and ID 02 with a few jokes. In the second half of the clip, ID 04 shows up unexpectedly with ID 01, ID 02, and ID 03. He has a look that is mostly joyful.

Fig. 7.4 illustrates how likely it is for someone to feel "happy." In the first part of the video sequence, ID 02 and ID 03 are able to portray a joyful countenance at times. After the first few seconds of the video, ID 02 doesn't seem to find his coworkers' jokes funny anymore. ID 04, who is in charge of the other three people, shows up in the middle of the incident and offers to share drinks with them. They have a joyous moment together for a few seconds (as seen by the steep spike in the center of Fig. 7.4) before getting down to business. The video makes it seem like ID 01 is a joyful person who smiles a lot.

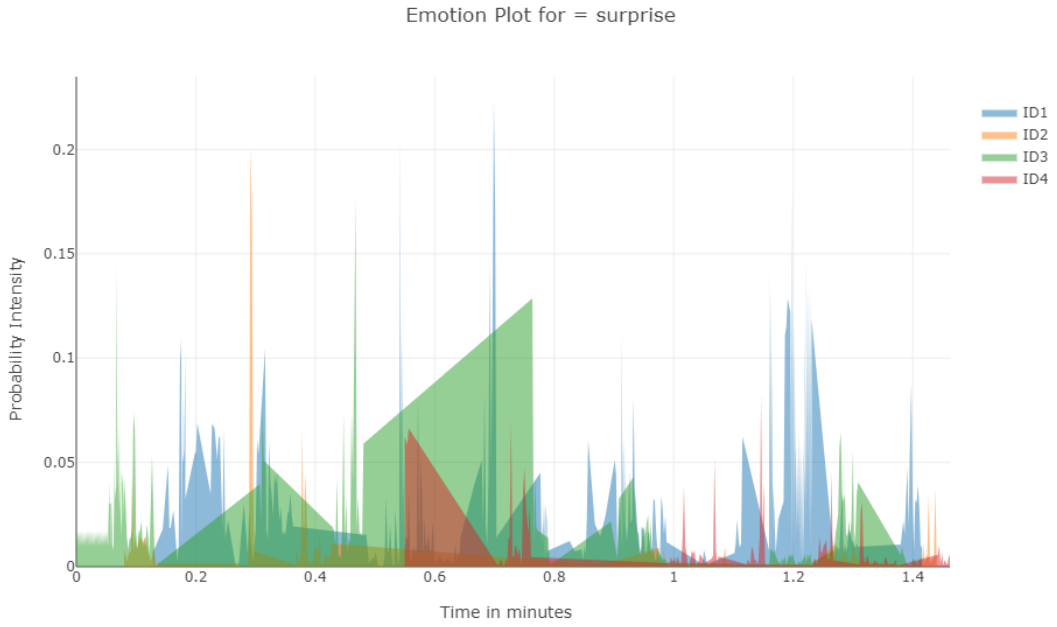


Figure 7.5: Area plot for emotion "Surprise"

Fig. 7.5 shows how the chances of the "surprise" emotion change. You can clearly see that ID 03 looks very surprised when ID 04 walks into the room. At that time, ID 04 also seems a little surprised. ID 02 shows practically no surprise at all in the whole sequence, which fits with how melancholy and silent he is in the film. ID 01 has had a face of astonishment for most of the scene. This is open to a more nuanced interpretation: ID 01 is the happiest of the four people in the video. Is it possible that the CNN can pick up on both happiness and surprise in some way?

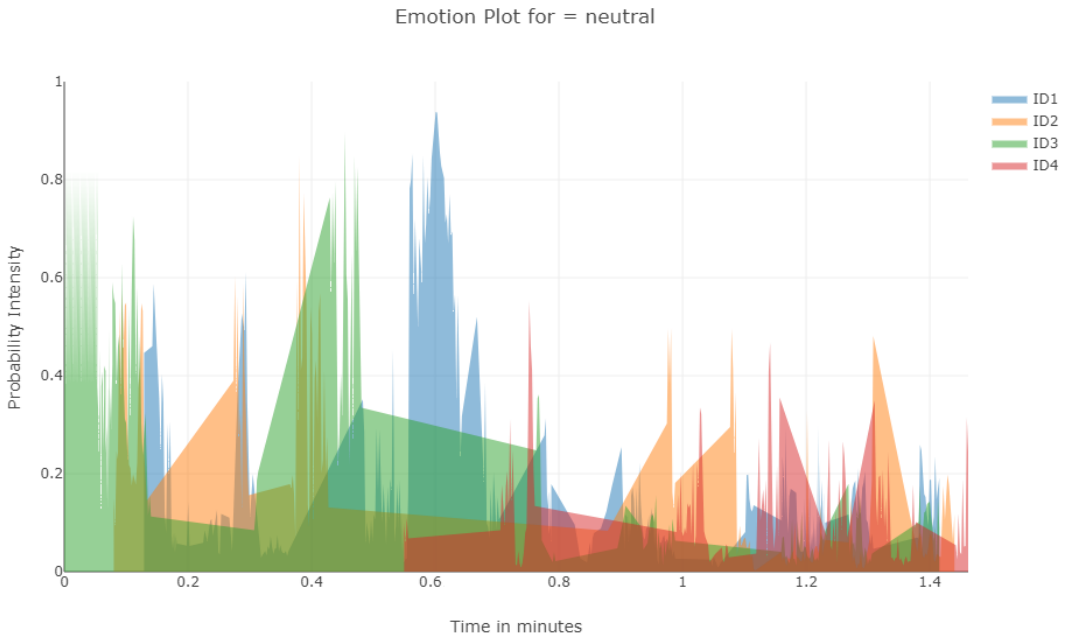


Figure 7.6: Area plot for emotion "Neutral"

Fig. 7.6 shows the different variations for the "neutral" emotion. Neutral qualities are a big part of the scene that stays the same. ID 03, who is believed to be the most professional and subtle of the four subjects, displays the most prominent neutral facial attributes. These traits fade as the video goes on, starting with ID 04. ID 02's neutral expressions stay the same the whole time. ID 04, who is higher up in the professional hierarchy, stays impartial as he enters the scene..

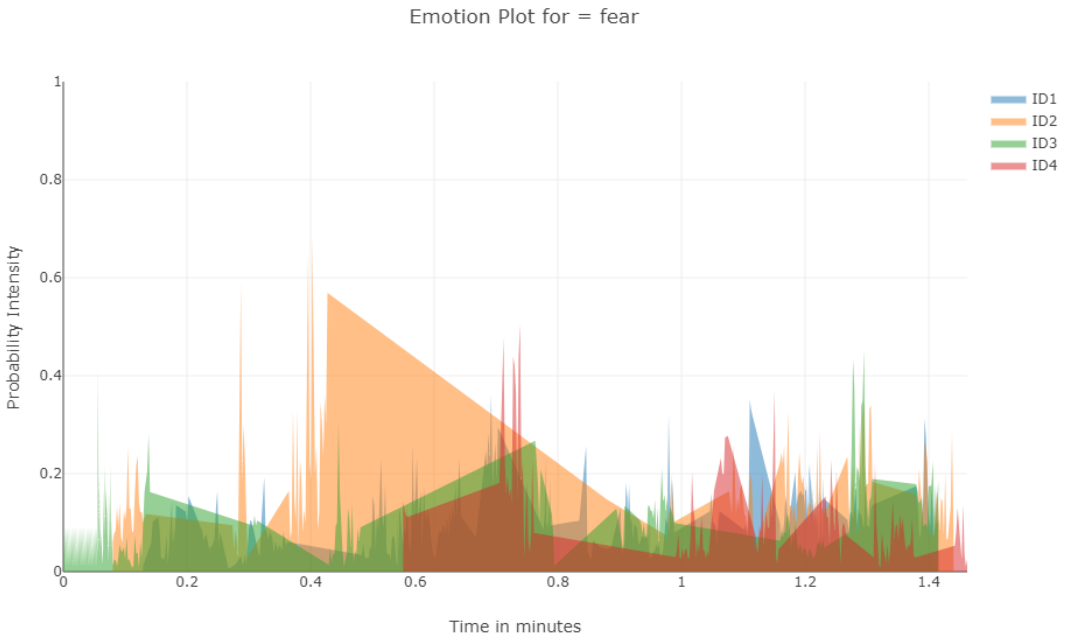


Figure 7.7: Area plot for emotion "Fear"

Fig. 7.7 shows the variations in the probability of the "Fear" emotion. Before ID 04 shows up, ID 02 shows a strong response to the fear trait. During this time, ID 02 was talking about the political problems they could have to deal with because of their recently finished mission. The graph suggests that ID 02 is giving the matter a lot of attention, which could explain why they were grumpy and didn't respond to the last three emotions. He keeps the fear feature in his expressions even when ID 04 comes into play, which demonstrates that he is really focused on the idea. ID 03 and ID 04 consistently display a low level of anxiety in their facial expressions. This could mean that they are more serious about their intentions and how they look as people. ID 01 is not really accountable for this class.

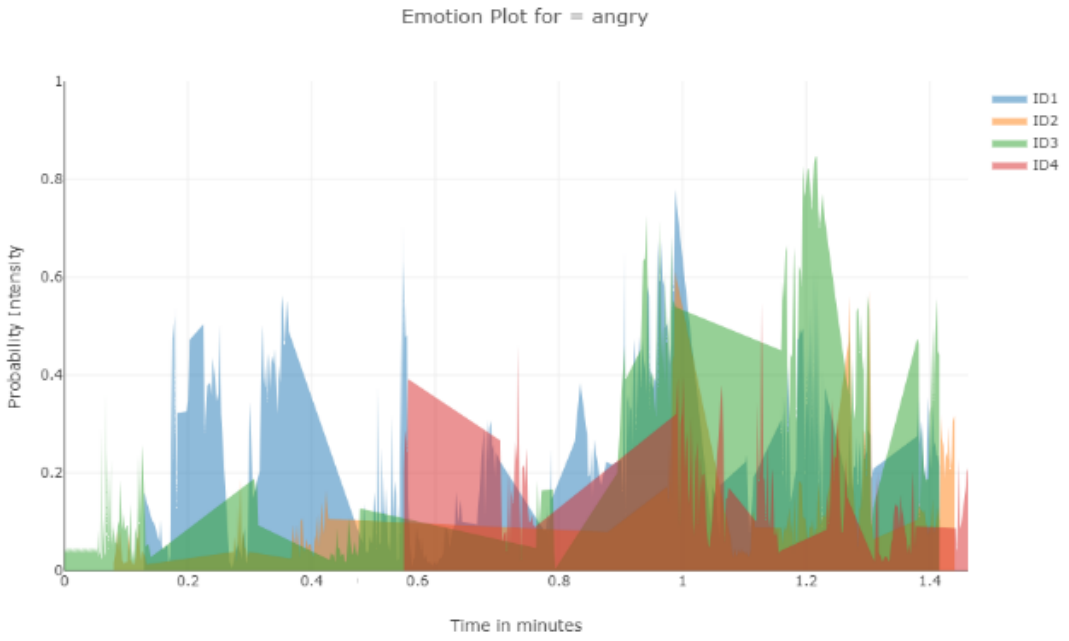


Figure 7.8: Area plot for emotion "Angry"

Fig. 7.8 shows the probability variations Regarding the "Angry" emotion, it responds particularly strongly to ID 01, ID 03, and ID 04. Now, the interpretation might be questioned because we have strong evidence for the emotional responses in the last four plots. Looking at the footage, it's not clear why the CNN saw signs of anger in the people in it. One simple argument is that the CNN misreads the hypothetical persons because of how their faces are shaped. The CNN has learned from real human faces, therefore it might have trouble with rendering the faces of fake video game characters.

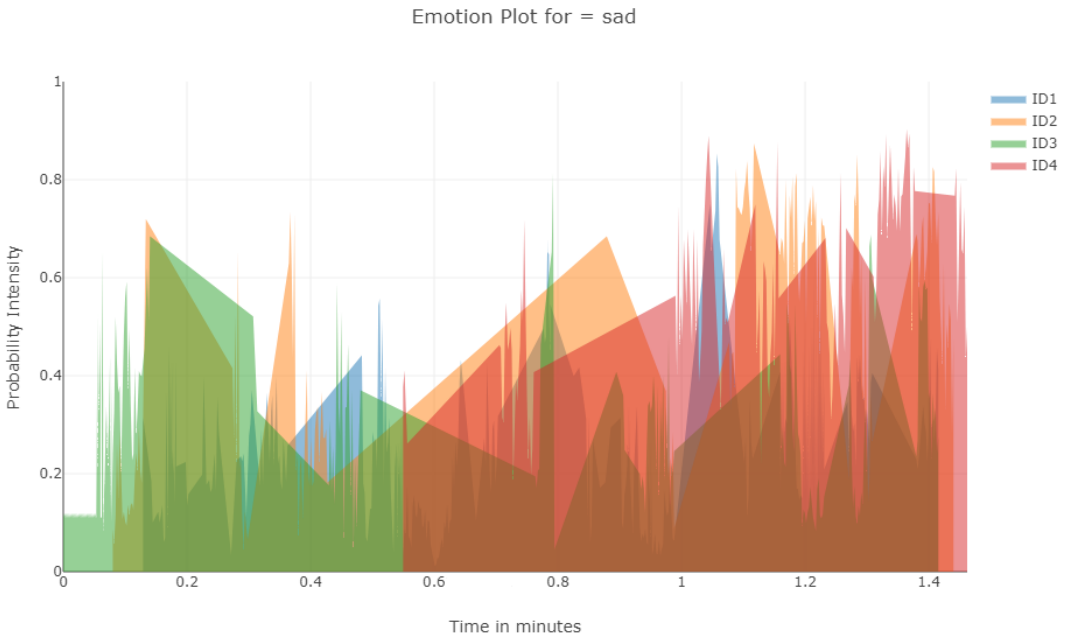


Figure 7.9: Area plot for emotion "Sad"

Fig. 7.9 shows the probability variations for the "Sad" feeling, and all of the people in the video had a very strong response, especially ID 03, ID 02, and ID 04. ID 01 does have certain times when it responds "sad," but not all the time. You can understand this prediction in the same way as the "neutral" face response. Since the scene in the video is sad and serious, everyone in it should have a serious look. Upon reviewing the video, it may be observed that time segments featuring an individual with a serious, solemn expression are where the model identified a significant degree of the "sadness" feature.

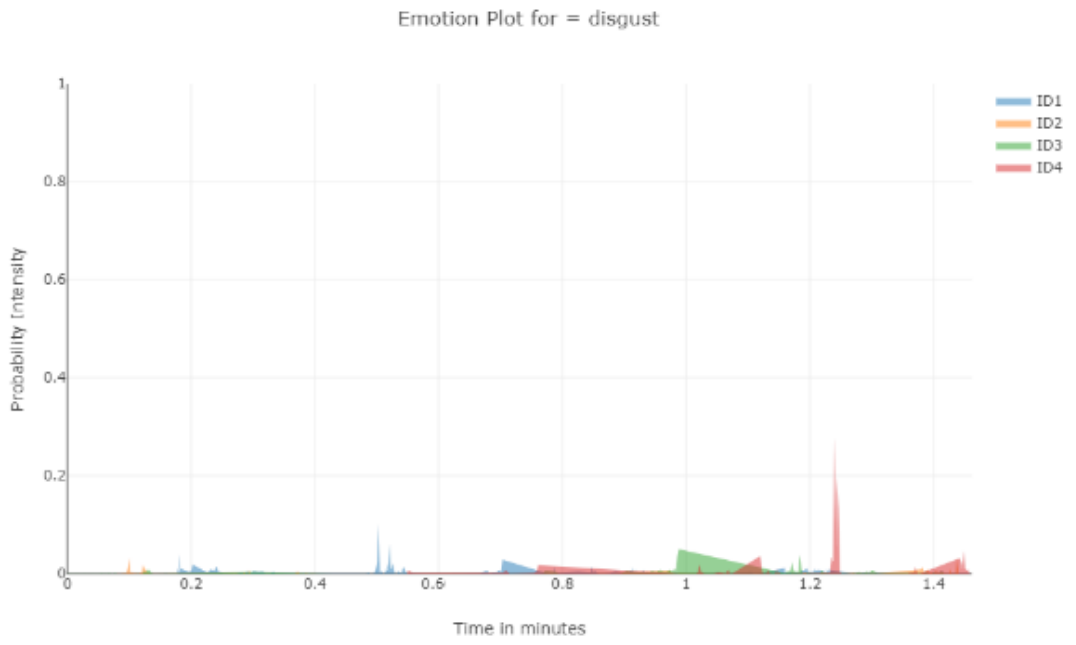


Figure 7.10: Area plot for emotion "Disgust"

Fig. 7.10 shows the probability variations for the sense of "Disgust." This feeling isn't in any aspect of the video, which is why the plot is like that. It is apparent that the probablity is less than 0.1

7.3 Discussion on Inherent Limitations

DNNs are rather limited in the kind of tasks they can execute well and the areas they can work in. Our experiment shows that the pretrained CNN for FER can only work with 7 emotion classes that it was trained on. Also, the softmax classification layer at the end can only give probability values, which means that the classes can't be the same. This is a fairly limited and constrained piece of information that doesn't let the HMI system make more complicated judgments later on. For instance, if the user is unsure or doesn't show any emotion on their face, the CNN might put the expression into one of the seven emotion classes. This would be a miserable guess to make future predictions on. It lacks the ability to respond, meaning it can't interpolate between the emotion classes it has been trained to predict for. Emotions are not mutually exclusive; individuals may exhibit expressions elicited by different emotions, such as angry and sadness.

The frame-based approach we used in our experiment is used by several video-based FER systems. This makes the engineering part easier and lets developers work in modules. The two CNNs, one for FER and the other for face detection, are trained and tested separately, and then they are combined to work together as

parts of a bigger pipeline. But this isn't the best technique to classify sequential data if the previous sample is linked to the next. Recurrent Neural Networks and its derivatives, such as LSTM and GRU, were specifically designed for deep learning with sequential input. In the context of our experiment's objective, the characteristics and prediction outputs are not well-defined. No research exists demonstrating that RNN versions can execute face detection or facial emotion recognition (FER). RNN variants are supervised algorithms that need the right annotations to tell them what to learn from the sequential data.

We were able to simply create a video-data-based FER system using our pipeline of two CNNs. When we do FER on a frame-by-frame basis, we get a set of emotion class prediction vectors for each face observed in each frame. Most videos have between 24 and 30 frames per second. The video we used for our experiment contained 25 frames per second. There were 3000 frames in all for the 2-minute clip. We got three prediction vectors of length 7 for each of the 3000 frames because IDs 01, 02, and 03 were available and detected virtually all of the time. Each prediction vector was related to a facial expression of a person that was taken at 1/25 of a second or 0.04 seconds. This length of time is too short for a human facial emotion to properly register, according to [174]. The forecasts we get for a particular subject wind up being very different from each other. CNNs don't always get their predictions right, and they might not always get the right expression. So, we got a bunch of predictions of face expressions taken every 0.04 seconds, but they might not be the same as the person's real facial expressions at all.

7.4 Summary

Our research presents two primary limitations of employing DNNs for FER, which hinder the capacity for HMI systems to achieve a more human-like inference of others' expressions and emotional states. We have talked about the challenge of classifying DL models assuming that classes are mutually exclusive. This kind of model can't make predictions about or fill in the gaps between the classes it was trained on. It is hard to fix the problem of large variance in frame-based forecasts. This is due to the absence of any modern work demonstrating the successful training of a FER-CNN with an RNN version.

The proposed remedies to the constraints outlined in the preceding section are predicated on an understanding of the mathematical model behind DNNs in general. Making changes to the structure to create new models that meet the needs of experimental goals can lead to much-needed progress in the field of HMI Systems. HMI systems that can recognize and keep track of how a person's

perceived emotional states change over time in a social group context are a viable target for employing machine learning in psychological, psychiatric, and human resource management studies.

CHAPTER 8

Conclusions, Limitations, and Future Research Directions

This thesis develops the study of ensemble learning for facial emotion recognition and proposes three independent yet complementary ensembles that synergistically improve the state of the art in affective computing. The contributions of the research and the novel methodologies consist of a holistic form enabling, the understanding and the use of ensemble learning concepts in the emotional recognition research area.

8.1 Theoretical innovations and approaches

There are many interlinked theoretical implications of our analysis. It is a big leap on the DNN architectures that work for the recognition of emotions by adding more activation functions in M3SI-Net. The above activation functions are designed according to the characteristics of subtle facial emotions. They illustrate the way of modifying base neural network elements to get a more efficient solution for a specific task. Information theoretical feature selection methods provide systematic mechanisms to identify important aspects of the emotional classification problem without calling for computationally expensive algorithms. This contribution is important, since the proposed theory reveals the properties that are better suited for emotion recognition, and how this knowledge can be used to create better systems. The alterations to the sigmoid-based ensemble topologies employed for SIG-Net suggest how, in a sense, we might learn how to make an ensemble combination technique perform better in certain domains. The findings are that ensemble methods to recognize emotions that take into account uncertainty and being able to provide reliable confidence measures are beneficial. That's important for users of the practical applications, because if the forecasts were wrong, they

could have big effects.

8.2 Experimental Analysis and Improvement of Performance

Validation experiment on various datasets and performance metrics demonstrate the superiority of multi-task ensembles over the state-of-the-art methods. The results contribute to the evidence on the superiorities of the new methods over the previous benchmark models. This study also demonstrates that the proposed approach can enhance not only the accuracy but also the reliability and robustness of the learned system. There's industry relevance here, too: Emotion detection systems are often used alongside autonomous systems at work in natural environments, and just being accurate isn't enough—but resilience is.The societal benefits of the proposed work extend across diverse domains, including healthcare application,educational technology,automotive safety,human-computer interaction and security .

8.3 Limitations

8.3.1 Dataset and evaluation limitation

The present study is an assessment of several datasets and settings, however, some limitations should be noted. Emotion detection data sets abound, but they don't always prove as effective with other cultures or individuals in the ways that they express their emotions. This limitation is specifically striking when we consider ensemble methods as they need many diverse training sets to generalize their models. The types of evaluation approaches were considered inclusive: however the authors' were constrained by the focus on emotional categories in, instead of continuous dimensional, models. Such a constraint may limit the generalizability of the results to other situations in which the ability to understand more complex emotions (e.g., in the context of therapy or education, where even small differences in emotional experiences are meaningful) is relevant.

8.3.2 Computational and scalability constraints

The model is computationally intensive, but the hardware and time in which to test is limited. This work also does not consider the long-term performance of algorithms (especially for ensemble techniques, that may decay at a different rate compared to single models) over an extended period of time. The scalability test

is what checks that; it looks mostly at how well the system puts up with extra computations, not extra data. With the emergence of new and larger emotion identification corpora, the scalability of ensembles could significantly change in ways that are poorly captured by previous studies.

8.3.3 Methodological diversity and transferability

The work is on identifying the emotions in faces. Although the powerful intuition underlying ensemble learning is expected to hold in general, the particular methodological developments might have to be adapted when considering another modality, like speech emotion recognition or multimodal emotion recognition. This is however a limitation of the present work, albeit one that paves the way for future studies and highlights the importance in specific cross-subject methodological scrutiny.

8.4 Future directions and further work

8.4.1 Multimodal Ensemble Learning Integration Proposal

It would be interesting to see if ensemble-learning methods could be employed for systems that recognize emotions in multiple ways as well. The approach developed in this dissertation provides a solid baseline for ensembles combining facial emotions with speech, physiology and environment. The problem to multimodal ensemble learning is, however, to not only organizing the many modalities, but also to model ethically devise techniques to deal with the different temporal properties, reliability distribution, and information gain mechanisms of each modality. Broadly, the information content principles we formulated in this study for modality selection could also be applied to the selection of the multimodal features and may even permit ensemble combination methods to work in a broader spectrum of multimodal systems.

8.4.2 Improving Ensembles Themselves

The findings of this study indicate that insight of such dynamic and experiential environments can be quite valuable, when developing adaptive ensemble systems that adjust their working to the evolving experiential contexts. We rather will concentrate overall performance by transitioning from static, uncertain configurations of an ensemble system (learning system) to more dynamic ones. Such systems would require some layer of high-level meta-learning that could rapidly learn exactly how healthy various parts of an ensemble were and how it could

condition their combination in various ways through intervention. This thesis builds a basis, namely the information-theoretic feature selection methods and the confidence calibration strategies of SIG-Net, for adaptive systems.

8.4.3 Privacy Preserving and Federated Learning

One of the novel research problems for the future will be how to employ ensemble learning methods in a federated learning setting when data privacy regulations prevent all of the data from being kept in a single location. Emotion recognition systems- particularly, in applications within healthcare, education, personal devices etc. need strong privacy protections, which cannot be provided by traditional ensemble learning methods. Work can further be put into designing ensemble learning algorithms for federated learning, where each local model learns on local data subsequently being bootstrapped using privacy-accessing protocols. There would be some methodological legwork involved, but this would allow emotion-identification systems to benefit from big datasets without violating users' privacy.

8.5 Conclusion

This thesis has demonstrated that ensemble learning is a powerful and versatile method for facial emotion recognition beyond mere accuracy improvement. The research contributes the theoretical basis, methodological innovation and practical framework, which jointly promote our understanding on how different models can be combined to build more robust, efficient and reliable emotion identification systems. The M3SI-Net, information-theoretic feature selection, and SIG-Net are the three primary methodological contributions. All of them address different facets of the ensemble learning problem and assists in constructing an overall view on the principles of ensemble learning. Chapters 3-7 in-depth evaluation template will allow us to select and apply the best ensemble methods according to the application characteristics and constraints. These findings suggest that the future of emotion recognition is not in training ever larger monolithic models, instead it lies in the careful selection and interlacing of different approaches which can adapt to changing requirements, give reliable performance across a wider set of conditions, and provide the required level of flexibility for deployment in real-world applications. The limitations identified include key directions for future research, and the broad relevance of ensemble learning principles suggests the methodological underpinnings outlined in this work will remain applicable as the field advances. As system of emotion recognition are being sought more and more in investment,

health care, education, gaming and human-computer interaction, the robust and reliable approaches developed in this thesis will be crucial to ensure that these systems work and are reliable. The transition from simple emotion expression models to more complex ensemble systems does indeed indicate a maturing of the field more in line with the general direction of artificial intelligence to both more reliable and explainable systems. This thesis contributes to the maturing of the field by providing theoretical underpinnings and practical solutions furthering the state-of-the-art as well as guiding principles for future developments in this central area of human-centered artificial intelligence.

References

[1] S. Tribedi and R. K. Barai, "M 3 si-net: A fusion model for facial emotion recognition with inception blocks and re-parameterized swish1 function," *Multimedia Tools and Applications*, pp. 1–24, 2025.

[2] S. Tribedi and R. K. Barai, "A lightweight deep feature selection contour for emotion recognition from human faces," *Journal of Signal Processing Systems*, pp. 1–12, 2025.

[3] S. Tribedi and R. K. Barai, "Limitations of facial emotion recognition using deep learning for intelligent human-machine interfaces," in *The Role of IoT and Blockchain*, Apple Academic Press, 2022, pp. 295–309.

[4] S. Tribedi and R. K. Barai, "Generating context-free group-level emotion landscapes using image processing and shallow convolutional neural networks," in *Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019*, Springer, 2020, pp. 313–325.

[5] D. Y. Liliana, "Emotion recognition from facial expression using deep convolutional neural network," in *Journal of physics: conference series*, IOP Publishing, vol. 1193, 2019, p. 012 004.

[6] J. W. Kusno and A. Chowanda, "Modeling emotion recognition system from facial images using convolutional neural networks," *CommIT (Communication and Information Technology) Journal*, vol. 18, no. 2, pp. 251–259, 2024.

[7] H. H. Chieng, N. Wahid, P. Ong, and S. R. K. Perla, "Flatten-t swish: A thresholded relu-swish-like activation function for deep learning," *arXiv preprint arXiv:1812.06247*, 2018.

[8] P. Ekman, W. V. Friesen, and S. S. Tomkins, "Facial affect scoring technique: A first validity study," 1971.

[9] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[10] L. A. Bartlett, S. Mawji, S. Whitehead, *et al.*, "Where giving birth is a forecast of death: Maternal mortality in four districts of afghanistan, 1999–2002," *The Lancet*, vol. 365, no. 9462, pp. 864–870, 2005.

[11] L. D. Cohn, "Sex differences in the course of personality development: A meta-analysis.," *Psychological bulletin*, vol. 109, no. 2, p. 252, 1991.

[12] C. Lyons, *Definiteness*. Cambridge University Press, 1999.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part I 6*, Springer, 2000, pp. 404–420.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] C.-H. Chou, S. Shrestha, C.-D. Yang, *et al.*, "Mirtarbase update 2018: A resource for experimentally validated microrna-target interactions," *Nucleic acids research*, vol. 46, no. D1, pp. D296–D302, 2018.

[18] A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[21] L. I. Kuncheva, "Classifier ensembles for changing environments," in *International workshop on multiple classifier systems*, Springer, 2004, pp. 1–15.

[22] L. Breiman, "Using iterated bagging to debias regressions," *Machine Learning*, vol. 45, pp. 261–277, 2001.

[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[24] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.

[26] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[27] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[28] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[29] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[30] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, IEEE, 2000, pp. 46–53.

[31] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[33] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[34] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete, "A survey on modern trainable activation functions," *Neural Networks*, vol. 138, pp. 14–32, 2021.

[35] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[36] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Computing and Applications*, pp. 1–18, 2021.

[37] Z. Wang, R. Jiao, and H. Jiang, "Emotion recognition using wt-svm in human-computer interaction," *Journal of New Media*, vol. 2, no. 3, p. 121, 2020.

[38] H. Kalantarian, K. Jedoui, P. Washington, *et al.*, "Labeling images with facial emotion and the potential for pediatric healthcare," *Artificial intelligence in medicine*, vol. 98, pp. 77–86, 2019.

[39] Z. Fei, E. Yang, D. D.-U. Li, *et al.*, "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212–227, 2020.

[40] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services," *Information Sciences*, vol. 479, pp. 416–431, 2019.

[41] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159 081–159 089, 2019.

[42] J. LE, "Some emotional aspects of prolonged illness in children.," *Public Health Nursing*, vol. 40, no. 5, pp. 257–260, 1948.

[43] G. REALE, "Considerations on facial expression in depression and schizophrenia," *Rassegna di Studi Psichiatrici*, vol. 39, no. 5-6, pp. 631–640, 1950.

[44] C. Darwin, *The expression of the emotions in man and animals, 1872*. 1872.

[45] P. Ekman, "Facial expressions," *Handbook of cognition and emotion*, vol. 16, no. 301, e320, 1999.

[46] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual review of psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[47] R. Ravi, S. Yadhukrishna, *et al.*, "A face expression recognition using cnn & lbp," in *2020 fourth international conference on computing methodologies and communication (ICCMC)*, IEEE, 2020, pp. 684–689.

[48] P. Lewinski, T. M. Den Uyl, and C. Butler, "Automated facial coding: Validation of basic emotions and facs aus in facereader.," *Journal of Neuroscience, Psychology, and Economics*, vol. 7, no. 4, p. 227, 2014.

[49] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[50] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, IEEE, 2016, pp. 1–10.

[51] X. Ben, Y. Ren, J. Zhang, *et al.*, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5826–5846, 2021.

[52] J. Cheng and G.-Y. Liu, "Affective detection based on an imbalanced fuzzy support vector machine," *Biomedical Signal Processing and Control*, vol. 18, pp. 118–126, 2015.

[53] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," *Affective computing, emotion modelling, synthesis and recognition*, vol. 2008, no. 5, pp. 421–440, 2008.

[54] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, Springer, 2012, pp. 808–822.

[55] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 3687–3691.

[56] M. Ranzato, V. Mnih, J. M. Susskind, and G. E. Hinton, "Modeling natural images using gated mrfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2206–2222, 2013.

[57] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, IEEE, 2014, pp. 211–214.

[58] X. Ben, C. Gong, T. Huang, C. Li, R. Yan, and Y. Li, "Tackling micro-expression data shortage via dataset alignment and active learning," *IEEE Transactions on Multimedia*, 2022.

[59] C. Xu, Y. Cui, Y. Zhang, P. Gao, and J. Xu, "Person-independent facial expression recognition method based on improved wasserstein generative adversarial networks in combination with identity aware," *Multimedia Systems*, vol. 26, pp. 53–61, 2020.

[60] Y. Kumar, S. K. Verma, and S. Sharma, "Multi-pose facial expression recognition using hybrid deep learning model with improved variant of gravitational search algorithm.," *Int. Arab J. Inf. Technol.*, vol. 19, no. 2, pp. 281–287, 2022.

[61] X. Zhu, S. Ye, L. Zhao, and Z. Dai, "Hybrid attention cascade network for facial expression recognition," *Sensors*, vol. 21, no. 6, p. 2003, 2021.

[62] B. Li, Y. Zhou, R. Xiao, *et al.*, "Unsupervised cross-database micro-expression recognition based on distribution adaptation," *Multimedia Systems*, vol. 28, no. 3, pp. 1099–1116, 2022.

[63] S. Xie and H. Hu, "Facial expression recognition with frr-cnn," *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.

[64] A. Agrawal and N. Mittal, "Using cnn for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy," *The Visual Computer*, vol. 36, no. 2, pp. 405–412, 2020.

[65] H. Li, X. Xiao, X. Liu, J. Guo, G. Wen, and P. Liang, "Heuristic objective for facial expression recognition," *The Visual Computer*, vol. 39, no. 10, pp. 4709–4720, 2023.

[66] Y. Xi, Q. Mao, and L. Zhou, "Weighted contrastive learning using pseudo labels for facial expression recognition," *The Visual Computer*, vol. 39, no. 10, pp. 5001–5012, 2023.

[67] G. Wen, H. Li, and D. Li, "An ensemble convolutional echo state networks for facial expression recognition," in *2015 international conference on affective computing and intelligent interaction (ACII)*, IEEE, 2015, pp. 873–878.

[68] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 435–442.

[69] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343–350, 2017.

[70] M. Sultan Zia, M. Hussain, and M. Arfan Jaffar, "A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier," *Multimedia Tools and Applications*, vol. 77, pp. 25 537–25 567, 2018.

[71] D. Ghimire and J. Lee, "Extreme learning machine ensemble using bagging for facial expression recognition," *Journal of Information Processing Systems*, vol. 10, no. 3, pp. 443–458, 2014.

[72] V. R. R. Chirra, S. R. Uyyala, and V. K. K. Kolli, "Virtual facial expression recognition using deep cnn with ensemble learning," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–19, 2021.

[73] D. Li, G. Wen, X. Li, and X. Cai, "Graph-based dynamic ensemble pruning for facial expression recognition," *Applied Intelligence*, vol. 49, pp. 3188–3206, 2019.

[74] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[75] D. Duncan, G. Shine, and C. English, "Facial emotion recognition in real time," *Computer Science*, pp. 1–7, 2016.

[76] H.-S. Lee and B.-Y. Kang, "Continuous emotion estimation of facial expressions on jaffe and ck+ datasets for human–robot interaction," *Intelligent service robotics*, vol. 13, pp. 15–27, 2020.

[77] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[78] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning," *Education and Information Technologies*, vol. 28, no. 4, pp. 4069–4092, 2023.

[79] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, IEEE, 2010, pp. 94–101.

[80] M. Mukhopadhyay, A. Dey, R. N. Shaw, and A. Ghosh, "Facial emotion recognition based on textural pattern and convolutional neural network," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 2021, pp. 1–6.

[81] N. Rathee, A. Vaish, and S. Gupta, "Emotion detection through fusion of complementary facial features," in *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, 2017, pp. 163–166.

[82] Y. Yu, H. Huo, and J. Liu, "Facial expression recognition based on multi-channel fusion and lightweight neural network," *Soft Computing*, pp. 1–15, 2023.

[83] I. Sobel, G. Feldman, *et al.*, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, pp. 271–272, 1968.

[84] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[85] A. Howard, M. Sandler, G. Chu, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[86] C. Han, Q. Wang, S. A. Dianat, *et al.*, "Amd: Automatic multi-step distillation of large-scale vision models," in *European Conference on Computer Vision*, Springer, 2024, pp. 431–450.

[87] V. Mayya, R. M. Pai, and M. M. Pai, "Automatic facial expression recognition using dcnn," *Procedia Computer Science*, vol. 93, pp. 453–461, 2016.

[88] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient facial expression recognition algorithm based on hierarchical deep neural network structure," *IEEE access*, vol. 7, pp. 41 273–41 285, 2019.

[89] R. I. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved facial expression recognition based on dwt feature for deep cnn," *Electronics*, vol. 8, no. 3, p. 324, 2019.

[90] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *The visual computer*, vol. 36, pp. 391–404, 2020.

[91] S. M. González-Lozoya, J. de la Calleja, L. Pellegrin, H. J. Escalante, M. A. Medina, and A. Benitez-Ruiz, "Recognition of facial expressions based on cnn features," *Multimedia tools and applications*, vol. 79, pp. 13 987–14 007, 2020.

[92] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[94] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[95] B. Yang, C. Zhu, F. W. Li, T. Wei, X. Liang, and Q. Wang, "Iaacs: Image aesthetic assessment through color composition and space formation," *Virtual Reality & Intelligent Hardware*, vol. 5, no. 1, pp. 42–56, 2023.

[96] N. Jiang, B. Sheng, P. Li, and T.-Y. Lee, "Photohelper: Portrait photographing guidance via deep feature retrieval and fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 2226–2238, 2022.

[97] X. Senhua, G. Liqing, W. Liang, and F. Wei, "Multi-scale context-aware network for continuous sign language recognition," *Virtual Reality & Intelligent Hardware*, vol. 6, no. 4, pp. 323–337, 2024.

[98] J. Zhu, Q. Zhang, L. Fei, *et al.*, "Fffn: Frame-by-frame feedback fusion network for video super-resolution," *IEEE Transactions on Multimedia*, vol. 25, pp. 6821–6835, 2022.

[99] A. H. Al-Jebrni, S. G. Ali, H. Li, *et al.*, "Sthy-net: A feature fusion-enhanced dense-branched modules network for small thyroid nodule classification from ultrasound images," *The Visual Computer*, vol. 39, no. 8, pp. 3675–3689, 2023.

[100] R. Liu, M. Liu, B. Sheng, *et al.*, "Nhbs-net: A feature fusion attention network for ultrasound neonatal hip bone segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3446–3458, 2021.

[101] R. Liu, T. Wang, H. Li, *et al.*, "Tmm-nets: Transferred multi-to mono-modal generation for lupus retinopathy diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1083–1094, 2022.

[102] Z. Cao, D. Liu, Q. Wang, and Y. Chen, "Towards unbiased label distribution learning for facial pose estimation using anisotropic spherical gaussian," in *European Conference on Computer Vision*, Springer, 2022, pp. 737–753.

[103] W. Wang, C. Han, T. Zhou, and D. Liu, "Visual recognition with deep nearest centroids," *arXiv preprint arXiv:2209.07383*, 2022.

[104] A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, "Comparing ensemble strategies for deep learning: An application to facial expression recognition," *Expert Systems with Applications*, vol. 136, pp. 1–11, 2019.

[105] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, "Facial emotion recognition using an ensemble of multi-level convolutional neural networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1 940 015, 2019.

[106] S. Narayan, "The generalized sigmoid activation function: Competitive supervised learning," *Information sciences*, vol. 99, no. 1-2, pp. 69–82, 1997.

[107] A. C. Marreiros, J. Daunizeau, S. J. Kiebel, and K. J. Friston, "Population dynamics: Variance and the sigmoid activation function," *Neuroimage*, vol. 42, no. 1, pp. 147–157, 2008.

[108] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[109] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[110] R. Kundu, H. Basak, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar, "Fuzzy rank-based fusion of cnn models using gompertz function for screening covid-19 ct-scans," *Scientific reports*, vol. 11, no. 1, p. 14 133, 2021.

[111] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[112] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

[113] A. Jamshidnejad and A. Jamshidined, "Facial emotion recognition for human computer interaction using a fuzzy model in the e-business," in *2009 Innovative Technologies in Intelligent Systems and Industrial Applications*, IEEE, 2009, pp. 202–204.

[114] M. Schröder, "Emotional speech synthesis: A review," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[115] S. Yang and B. Bhanu, "Understanding discrete facial expressions in video using an emotion avatar image," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 980–992, 2012.

[116] X. Ji, H. Zhou, K. Wang, *et al.*, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[117] K. Wang, Q. Wu, L. Song, *et al.*, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*, Springer, 2020, pp. 700–717.

[118] N. Kumari and R. Bhatia, "Efficient facial emotion recognition model using deep convolutional neural network and modified joint trilateral filter," *Soft Computing*, vol. 26, no. 16, pp. 7817–7830, 2022.

[119] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimedia Tools and Applications*, pp. 1–35, 2022.

[120] 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.

[121] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*, IEEE, 1998, pp. 200–205.

[122] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, IEEE, 2010, pp. 94–101.

[123] I. J. Goodfellow, D. Erhan, P. L. Carrier, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer, 2013, pp. 117–124.

[124] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[125] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[126] J. Irvin, P. Rajpurkar, M. Ko, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 590–597.

[127] J. N. Kather, J. Krisam, P. Charoentong, *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, e1002730, 2019.

[128] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015.

[129] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[130] S. Chattopadhyay, P. K. Singh, M. F. Ijaz, S. Kim, and R. Sarkar, "Snapensemfs: A snapshot ensembling-based deep feature selection model for colorectal cancer histological analysis," *Scientific Reports*, vol. 13, no. 1, p. 9937, 2023.

[131] M. Ghosh, R. Guha, R. Mondal, P. K. Singh, R. Sarkar, and M. Nasipuri, "Feature selection using histogram-based multi-objective ga for handwritten devanagari numeral recognition," pp. 471–479, 2018.

[132] M. Ghosh, T. Kundu, D. Ghosh, and R. Sarkar, "Feature selection for facial emotion recognition using late hill-climbing based memetic algorithm," *Multimedia Tools and Applications*, vol. 78, pp. 25 753–25 779, 2019.

[133] A. Marik, S. Chattopadhyay, and P. K. Singh, "A hybrid deep feature selection framework for emotion recognition from human speeches," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11 461–11 487, 2023.

[134] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[135] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, e87357, 2014.

[136] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, 1992.

[137] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[138] M. Sajid, N. Iqbal Ratyal, N. Ali, *et al.*, "The impact of asymmetric left and asymmetric right face images on accurate age estimation," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[139] N. I. Ratyal, I. A. Taj, M. Sajid, N. Ali, A. Mahmood, and S. Razzaq, "Three-dimensional face recognition using variance-based registration and subject-specific descriptors," *International Journal of Advanced Robotic Systems*, vol. 16, no. 3, p. 1 729 881 419 851 716, 2019.

[140] S. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *2012 4th International conference on intelligent human computer interaction (IHCI)*, IEEE, 2012, pp. 1–5.

[141] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimedia Tools and Applications*, vol. 76, pp. 7803–7821, 2017.

[142] C. Benitez-Quiroz and R. Srinivasan, "&Amp;amp; martinez, am (2016). emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5562–5570.

[143] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern recognition*, vol. 61, pp. 610–628, 2017.

[144] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization cnn," in *2018 37th Chinese control conference (CCC)*, IEEE, 2018, pp. 9608–9613.

[145] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," pp. 118–126, 2017.

[146] S. Shaees, H. Naeem, M. Arslan, M. R. Naeem, S. H. Ali, and H. Aldabbas, "Facial emotion recognition using transfer learning," pp. 1–5, 2020.

[147] S. Malakar, M. Ghosh, S. Bhowmik, R. Sarkar, and M. Nasipuri, "A ga based hierarchical feature selection approach for handwritten word recognition," *Neural Computing and Applications*, vol. 32, pp. 2533–2552, 2020.

[148] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A micro-ga embedded pso feature selection approach to intelligent facial emotion recognition," *IEEE transactions on cybernetics*, vol. 47, no. 6, pp. 1496–1509, 2016.

[149] S. Saha, M. Ghosh, S. Ghosh, *et al.*, "Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm," *Applied Sciences*, vol. 10, no. 8, 2020, ISSN: 2076-3417. DOI: 10.3390/app10082816.

[150] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2017, pp. 4700–4708.

[151] A. Paszke, S. Gross, S. Chintala, *et al.*, "Automatic differentiation in pytorch," 2017.

[152] *Kaggle: Your machine learning and data science community*, https://www.kaggle.com.

[153] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[154] H. Zhang, B. Huang, and G. Tian, "Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture," *Pattern Recognition Letters*, vol. 131, pp. 128–134, 2020.

[155] Y. Tang, X. Zhang, X. Hu, S. Wang, and H. Wang, "Facial expression recognition using frequency neural network," *IEEE Transactions on Image Processing*, vol. 30, pp. 444–457, 2020.

[156] T. Kalsum, Z. Mehmood, *et al.*, "A novel lightweight deep convolutional neural network model for human emotions recognition in diverse environments," *Journal of Sensors*, vol. 2023, 2023.

[157] P. Foggia, A. Greco, A. Saggese, and M. Vento, "Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105 651, 2023.

[158] A. P. Fard and M. H. Mahoor, "Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26 756–26 768, 2022.

[159] A. Sultana, S. K. Dey, and M. A. Rahman, "Facial emotion recognition based on deep transfer learning approach," *Multimedia Tools and Applications*, pp. 1–15, 2023.

[160] K. N. Kumar Tataji, M. N. Kartheek, and M. V. Prasad, "Cc-cnn: A cross connected convolutional neural network using feature level fusion for facial expression recognition," *Multimedia Tools and Applications*, pp. 1–27, 2023.

[161] M. Lyons, M. Kamachi, and J. Gyoba, "The japanese female facial expression (jaffe) dataset," *(No Title)*, 1998.

[162] J. Cohn, "Kanade. cohn-kanade au-coded facial expression database," Technical report, Pittsburgh University, Tech. Rep., 1999.

[163] M. Gevorgyan, A. Mamikonyan, and M. Beyeler, *OpenCV 4 with Python Blueprints: Build creative computer vision projects with the latest version of OpenCV 4 and Python 3*. Packt Publishing Ltd, 2020.

[164] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, Ieee, vol. 1, 2001, pp. I–I.

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

[166] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and vision computing*, vol. 47, pp. 3–18, 2016.

[167] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, Springer, 2012, pp. 679–692.

[168] B. C. Csáji *et al.*, "Approximation with artificial neural networks," *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24, no. 48, p. 7, 2001.

[169] A. Geitgey, "Face recognition documentation," *Release*, vol. 1, no. 3, pp. 3–37, 2019.

[170] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[171] B. E. Santoso and G. P. Kusuma, "Facial emotion recognition on fer2013 using vggspinalnet," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 7, pp. 2088–2102, 2022.

[172] N. Pavlichenko and D. Ustalov, "Imdb-wiki-sbs: An evaluation dataset for crowdsourced pairwise comparisons," *arXiv preprint arXiv:2110.14990*, 2021.

[173] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[174] S. L. Fayolle and S. Droit-Volet, "Time perception and dynamics of facial expressions of emotions," *PLoS One*, vol. 9, no. 5, e97944, 2014.

Signature of the Candidate : Sabyasachi Tripathi

(Author's Name)

Date : 10/11/25

10.11.2025

Professor
Electrical Engineering Department
JADAVPUR UNIVERSITY
Kolkata - 700 032