Multi Feature Fusion based Facial Expression Recognition using Deep Learning Techniques

Submitted in partial fulfilment of the requirements

for the award of the degree of

Doctor of Philosophy

by

Alagesan Bhuvaneswari Ahadit

(Roll No: 718046)

Under the supervision of

Prof. J. Ravi Kumar

Department of Electronics & Communication Engineering

National Institute of Technology Warangal

Telangana, India - 506004

2022

Dedicated

To

My Family,

Teachers & Friends

Declaration

This is to certify that the work presented in this thesis entitled **Multi Feature Fusion based Facial Expression Recognition using Deep Learning Techniques** is a bonafied work done by me under the supervision of **Prof. J. Ravi Kumar** and was not submitted elsewhere for the award of any degree.

I declare that this written submission represents my own ideas and even considered others ideas which are adequately cited and further referenced the original sources. I understand that any violation of the above will cause disciplinary action by the institute and can also evoke panel action from the sources or from whom proper permission has not been taken when needed. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea or data or fact or source in my submission.

Place: Warangal
Date: 17/8/2022

Alagesan Bhuvaneswari Ahadit
Research Scholar
Roll No.: 718046

NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL, INDIA-506004

Department of Electronics & Communication Engineering

CERTIFICATE

This is to certify that the thesis work entitled **Multi Feature Fusion based Facial Expression Recognition using Deep Learning Techniques** is a bonafide record of work carried out by **Alagesan Bhuvaneswari Ahadit** submitted to the faculty of **Electronics & Communication Engineering** department, in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy** in **Electronics and Communication Engineering, National Institute of Technology Warangal, India-506004**. The contributions embodied in this thesis have not been submitted to any other university or institute for the award of any degree.

Prof. J. Ravi Kumar
Research Supervisor
Professor
Department of ECE
NIT Warangal, India-506 004.

Place: Warangal
Date: 17/8/2022

Acknowledgements

First, I take immense pleasure to convey my sincere gratitude to my supervisor Prof. J. Ravi Kumar for his perpetual encouragement and supervision. His guidance has oriented me in a proper direction and supported me with promptness and care.

I take this privilege to thank all my Doctoral Scrutiny Committee members and Faculty members Prof. P. Sreehari Rao, Prof. L. Anjaneyulu, Prof. N. Bheema Rao, Dr. P. Muralidhar, Dr. Maheshwaram Satish, Dr. Raju Bhukya (Dept. of CSE), and Dr. M. Srinivas (Dept. of CSE), and for their detailed review, constructive suggestions, and excellent advice during the progress of this research work. I would also like to thank all the faculty of Dept. of ECE who helped me during the course.

I would also like to extend my heartfelt appreciation to my family, colleague scholars, friends, and well-wishers who helped to write my thesis with their support. Finally, I thank my nation India, for giving me the opportunity to carry out my research work at the NIT Warangal. A special thanks to MHRD for its financial support.

Ahadit A.B

Abstract

Automatic facial expression recognition (FER) has been a key task in cognitive science, machine intelligence, and computer vision since facial expressions are a common technique to assess human emotions. Research into automatic facial expression recognition (FER) models has been carried out extensively in recent years. In human–computer interaction (HCI), security monitoring, sociable robots, advanced driver assistant systems (ADASs), clinical psychology, and emotion analysis, functional applications of FERs are used. The FER models typically use hand-engineered techniques such as local binary patterns (LBP), non-negative matrix factorization (NMF), and scale-invariant feature transform (SIFT). The extracted characteristics are then given to a machine learning classifier to understand the patterns concealed in the features. Support vector machine (SVM), Ad-boost, and Random Forest are the standard machine learning models used for the classification of expressions in traditional FERs. The downside of handcrafted characteristics is that the classification task cannot be generalized because the algorithms adopt unique human-designed learning styles. The precision can be influenced by the reliance on geometry and the type of dataset. Researchers have switched to deep learning-based FER models to address these disadvantages. In this method, models are allowed to evaluate the patterns independently where models can extract low and high-level features on different face images through deep-linked convolutional neural layers. However, there are many critical issues in the field of facial expression recognition which should be addressed. The accuracy of the FER model is reduced due to problems such as the variations in expressing emotions, variations in lighting, and different ethnic biases. The taken research work solves these critical issues to improve the efficiency of the FER models.

There exist multiple challenges in designing an accurate robust FER model. The critical issue with the design of the FER model is the strong intra-class correlation of different emotion classes. It is a challenging task to classify accurate facial expressions due to

high intra-class correlation. The latest convolutional neural network-based FER models have shown significant improvement in accuracy score but lack distinguishing the micro-expressions. Although it has improved the accuracy, FER models must consider problems like ethnic bias and illumination variance. The CNN models also experience over-fitting problem when trained with sparse and class imbalanced samples. The discriminative ability of the CNN models which are used in general object classification tasks is not sufficient for facial expression classification since there exists high similarity between various emotion (label) classes. So, it is essential to design or modify existing FER model that helps in overcoming these issues. The related works shows that the multi feature fusion techniques could improve the accuracy of FER models. The next important problem that arises is the selection of features. There exist various computer vision algorithms to extract object features and it is essential to investigate which features could better represent facial expressions. These problems in the field of facial expression recognition are studied in this research work. The research work majorly focuses on improving the classification accuracy of FER models by considering each drawback of existing models.

Contents

List of Figures

List of Tables

List of Abbreviations

FER	Facial Expression Recognition
CNN	Convolutional Neural Network
HCI	Human-Computer Interaction
ADAS	Advanced Driver Assistant Systems
AR	Augment Reality
VR	Virtual Reality
LBP	Local Binary Patterns
NMF	Non-Negative Matrix Factorisation
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
DNN	Deep Neural Network
FACS	Facial Action Coding Systems
ViS	Uncontrolled Visible Light
NIR	Near Infrared Region
CNN-RNN	Convolutional And Recurrent Neural Networks
GANs	Generative Adversarial Networks
HOG	Histogram of Oriented Gradients
VGG-Face	Visual Geometry Group- Face
CK+	Extended Cohn-Kanade Dataset
KDEF	Karolinska Directed Emotional Faces
AUC	Area Under the Curve
RoC	Rank of Confidence
RMSProp	Root Mean Square Propagation
AdaGrad	Adaptive Gradient Algorithm
ADAM	Adaptive Moment Estimation

Chapter 1

Introduction

1.1 Introduction

Facial expression recognition (FER) is the process of identifying human emotions from facial expressions and facial expression detection also observes and analyzes human behavior traits. Facial expressions play a vital role in human nonverbal communication, and it helps to understand the inner feelings of humans. Human emotion analysis requires automated Facial Expression Recognition (FER) of unique facial features. According to research, nonverbal communication may represent nearly 55% of information during the interaction of humans [8]. Facial expressions can be voluntary or involuntary actions that can be usually observed with a naked eye. At times, few expressions may not be visible to the naked eye. Hence, there is a challenge to identify emotions automatically. There is much evidence that a few facial expressions can be mapped to a particular emotion like a smiling expression can be related to an emotional state of happiness [9]. Humans have the instinctive, natural ability to comprehend the emotion of a person just by observing facial expressions. There is a rapid attraction in the research of automatic facial emotion recognition in recent years. The applications of this topic include, but are not limited to, human-computer interaction (HCI), security monitoring, sociable robots, advanced driver assistant systems (ADASs), Augment reality (AR), Virtual reality (VR), Psychiatry, Pain assessment, Lie detection clinical psychology, sentiment analysis, and the entertainment business. Ekman and Friesen proposed six basic emotions related to cross-cultural studies [10]. The essential facial emotions discussed are Anger, Disgust,

Fear, Happiness, Sadness, and Surprise. There are other means of identifying human emotions through speech, text, and other biomedical data such as EEG [11]. Emotion recognition through facial features is simple as it does not require sophisticated sensors or transducers for extracting information.

Sensor data such as Electromyography (EMG) and Electrocardiogram (ECG) can also predict emotions. However, because of its simplicity, the camera is a favoured sensor as it is not necessary to be connected to humans. Automatic FER models can be categorised generally as two methods, traditional computer vision models and deep learning-based FER models. In order to extract features in the images, vision-based models typically use hand-engineered techniques such as local binary patterns (LBP), non-negative matrix factorization (NMF), and Scale-Invariant Feature Transform (SIFT). The extracted characteristics are then given to a machine learning classifier to understand the patterns concealed in the features. Support Vector Machine (SVM), Adaboost, and Random Forest are the standard machine learning models used in traditional FERs. In detecting emotions, this approach has achieved good accuracy. The downside of handcrafted characteristics is that the classification task cannot be generalised because the algorithms adopt unique human-designed learning styles. The precision can be influenced by the reliance on geometry and the type of dataset. Researchers have switched to deep learning-based FER models to address these disadvantages. In this method, models are allowed to evaluate the patterns independently. The models can extract low and high-level features on different face images through deep-linked convolutional neural layers. The main characteristic of deep learning-based FERs is that to get well trained in classifying emotions, the models need to analyse massive number of face images. The use of GPU becomes mandatory due to the high computation seen in deep learning models such as convolutional neural networks and recurrent neural networks. Deep learning frameworks provide building blocks for developing, training, and evaluating deep neural networks through a high-level programming interface. GPU-accelerated libraries like NVIDIA CUDA® Deep Neural Network library (cuDNN) are utilized by popular deep learning frameworks like TensorFlow, PyTorch, and others to offer high-performance multi-GPU accelerated training. The models designed in this research work use Keras (running on top of TensorFlow) deep learning API to build the various multi-feature fusion architectures. The Tensorflow GPU-supported libraries recommend NVIDIA® GPU cards that support CUDA® ar-

chitectures. However, the detailed requirement of the GPU specifications like Core Count, Core Clock Speed, Memory size, type, and bandwidth depends on the application and model complexity.

1.2 Issues in Facial Expression Recognition Models

The deep neural networks employed in FER have several issues to be tackled and in the research work, we have considered some of the critical issues that need to be handled to improve the efficiency and robustness of the existing deep neural network (DNN) based FER models. The following issues below are considered in the existing DNN-based models.

- The existing FER methods are not able to significantly overcome the intra-class correlation effect between various emotion classes.
- The available FER techniques should address the issue of overfitting during training of deep neural networks with sparse and class imbalanced samples.
- The FER models fail to deal with illumination variance issue. The models also show different accuracy levels in different lighting conditions. So the models should impart illumination invariance property in the design
- There exists only few FER models that have explored the efficiency of feature fusion DNNs. There is much scope in exploring and fusing various various handcrafted and deep learnt features.
- There exists no significant metrics which can measure the credibility of the model's classification. The existing probability scores calculated for a given prediction by DNN models are insufficient to measure the credibility of prediction.

The first issue is the intraclass correlation where certain emotions are strongly connected to other emotions since they share common micro expressions which are triggered during the emotion. The emotion of fear is strongly connected to the emotion surprise since there exist similar action units that are triggered in both emotions. The high misclassification rate observed in the FER models is due to the strong intra-class correlation

of emotions. Due to the similar emotion classes the FER models tend to get confused in classifying the emotions and this is where the traditional single convolutional neural network (CNN) based FER models usually fail. The detailed study of action units and the correlation of different emotions are studied in this research. The chapter 3 does the exploratory data analysis of various action units spatially on human face based on facial action coding systems (FACS). The research work in The first issue is the intraclass correlation where certain emotions are strongly connected to other emotions since they share common micro expressions which are trigged during the emotion. The emotion of fear is strongly connected to the emotion surprise since there exist similar action units that are triggered in both emotions. The high misclassification rate observed in the FER models is due to the strong intra-class correlation of emotions. Due to the similar emotion classes the FER models tend to get confused in classifying the emotions and this is where the traditional single convolutional neural network (CNN) based FER models usually fail. The detailed study of action units and the correlation of different emotions are studied in this research. The chapter 3 does the exploratory data analysis of various action units spatially on human face based on facial action coding systems (FACS). The research work in chapter 3 discusses the amount of correlation between each emotion class and also gives solution on how to avoid this intra class correlation effect in classifying the emotions by introducing a novel LogicMax layer. The chapter 3 explains different techniques to counter the intra-class correlation effect in discriminating the facial expressions.

The next issue discussed in this research work is regarding the overfitting effect seen in training the deep neural networks. The problem with the deep neural network is that efficient training of models requires huge datasets. Even after the intensive training few models tend to overfit to the provided datasets and do not generalize to the external data. The research takes this issue into consideration. The issue of relying only on deep features can sometimes cause issue in learning. The recent techniques like attention models can counter this effect. But instead of relying only one deep feature we have constructed a multi feature fusion model. In the experimentation done in the research work, we have observed that training the DNN with multi features can make models robust and learn faster when compared to conventional CNN related models. The selection of features in fusing and type of fusing is discussed in the chapter 4. The chapter 4 discusses existing multi feature fusion models and compares the efficiency of different features in

understanding the micro expressions of face. The detailed study of different metrics of hybrid model are compared with existing FER models. The chapter 4 also compares the prediction capability of proposed multi feature fusion model with existing FER models to describe the robustness of the proposed model.

The measure of credibility of various ML and DL models are discussed in the chapter 5. The section also explains why we need to measure the credibility of models. In general there exists many FER models that can classify facial expressions but there happens a dilemma in selection of the accurate model. The credibility of prediction needs to be measured certainly in critical conditions. The existing classification models anyhow describes the confidence score of prediction but deciding only on the confidence score can be a problem since there exists many false positives and false negatives of higher confidence scores. So, in the chapter 5 we have proposed a credibility score metric called "Rank of Confidence" (RoC). SternNet, a novel classification model is introduced in this research work. SternNet is a multi-stage classification model which classifies samples based on Rank of Confidence metrics. In the chapter 5 we discuss how the RoC is calculated and the discuss the overall flow process of SternNet.

The final issue in FER that is considered in the research work is the model reliance on illumination variations. The variations in illumination can seriously degrade the performance of FER models. The effect of illumination variance is serious issue since certain models are trained only on specific illumination (strong/dark/weak) lighting condition. We need to impart the illumination invariance property into our FER models. In certain critical conditions there is a need to run DL and ML models even in presence of low light conditions. So, the proposed model considers Near Infrared region based FER model in designing illumination invariant classification. The proposed model uses both spatial and temporal information for predicting the emotion class in near infra-red region (NIR).

1.3 Organization of Thesis

The organization of the research work discussed in this thesis can be described as follows:

- In section 2, the related works discuss the existing works of existing FER models. The section describes the methodology and brief explanation regarding the concepts of the design of FER models. The essential brief literature review is done in this section to understand the existing FER models.

- In section 3, we discuss the existing works of Facial action unit coding (FACS). The section describes how FACS information is used in designing a priority table using LogicMax layer. The section explains how LogicMax layer helps in discriminating highly correlated features.

- The section 4 describes the importance of HOG and VGG-Face features in modeling the Facial features. The section also discusses the detailed procedure of extracting HOG and CNN features on faces. The visualization and advantages of fusing the VGG-Face and HOG features are also explained in this section. The advantages of multi-feature fusion models over the traditional models are compared with various metrics like accuracy, precision, recall, and F1-score.

- The section 5, explains the architecture of the SternNet. The SternNet is a classification model which relies on stern rules in selecting the accurate class. The proposed model considers Rank of Confidence (RoC) which is credibility measuring metrics in classification problems.

- The section 6, demonstrates the fusing of spatial and temporal features in designing a CNN-RNN based deep neural network. The proposed CNN-RNN model is built in classifying expressions in Near-infrared videos. The NIR-based images/ videos can display faces even in low-light regions. The proposed hybrid model is tested on three different illumination conditions (strong/weak and dark).

- The section 7, demonstrates the cross-validation results of three datasets using the proposed multi-feature fusion and conventional CNN models. The comparison of confusion matrices and other classification metrics is discussed in this part. The visualization of prediction variations of CNN and the proposed model is also discussed in this section.

Chapter 2

Literature Review

Pre-processing, deep feature learning, and deep feature classification are the three primary processes that are frequent in automatic deep FER, and they are described in this section. The general preprocessing techniques involve Face alignment, Data augmentation, Face normalization, Illumination normalization, and Pose normalization. After identifying the face using a set of training data, background and non-facial areas are subsequently eliminated. The Viola-Jones face detector is a well-known and commonly used face detection implementation. It is reliable and computationally straightforward to detect near-frontal faces. Even while face detection is the sole process that is necessary to enable feature learning, additional face alignment that uses the coordinates of nearby landmarks can significantly improve FER performance because it can lessen the difference in face scale and in-plane rotation, this step is quite important. The process of facial expression recognition pipeline of FER models is shown in the Fig. 2.2

For deep neural networks to be generalizable to a specific recognition task, there must be enough training data. However, there aren't enough photos in the majority of publicly accessible FER databases to use for training. Data augmentation is therefore an essential stage in deep FER. On-the-fly data augmentation and offline data augmentation are the two categories into which data augmentation techniques can be categorised. Deep learning toolkits frequently include on-the-fly data augmentation to reduce overfitting. The input samples are randomly selected from the four corners and the center of the image during the training stage, and they are then flipped horizontally, which can produce a dataset that is 10 times bigger than the initial training data. Various offline data

augmentation techniques have been developed in addition to the fundamental on-the-fly data augmentation to further increase data on both size and diversity. The random perturbations and transforms, such as rotation, shifting, skew, scaling, noise, contrast, and colour jittering are the most often employed operations. For instance, to increase the amount of data, typical noise models like salt-and-pepper and speckle noise and Gaussian noise are used. Additionally, each pixel's saturation and value (the S and V components of the HSV colour space) are altered for contrast transformation to enrich the data. Combining many processes can produce more unused training samples and increase the network's resistance to rotated and deviated faces. The important data augmentation techniques implemented by the DNN models are shown in the Fig. 2.1

Normalization of illumination: In unrestricted situations, illumination and contrast can differ in different photographs even from the same individual with the same expression, which can lead to significant intraclass variations. In, three commonly used illumination normalization techniques, including difference of Gaussian (DoG), discrete cosine transform (DCT), and isotropic diffusion (IS)-based normalization, were assessed for illumination normalization. and removed illumination normalization using homomorphic filtering-based normalizing, which has been shown to produce the most reliable results out of all the other procedures. Additionally, comparable research has demonstrated that histogram equalization in combination with lighting normalization enhances face recognition performance over illumination normalization alone.

Paul Ekman [10] has identified anger, disgust, fear, happiness, sadness, and surprise as six basic emotions, later neutral was added to the list. Ekman introduced FACS [7], which is one of the iconic works made in the field of facial emotion recognition, helped many researchers to extend the work in this field. There are numerous works in the field of facial expression recognition. The architectures used in this field can be broadly classified into the following categories

1. Pretraining and fine-tuning based Neural Networks.
2. Multiple feature input networks.
3. Spare blocks and layers based deep neural networks.
4. Ensemble-based deep neural networks.

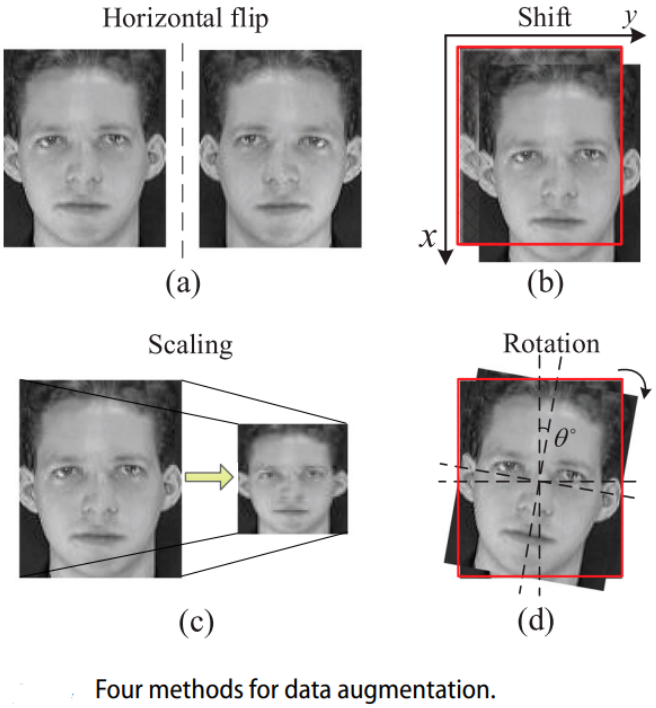

Figure 2.1 General data augmentation in four methods [2]

5. Generative adversarial networks based FER models

Pretraining and fine-tuning based Neural Networks use pre-trained networks like AlexNet [12], VGG [13], VGG-face [14], and GoogleNet [15]. The motivation behind using these pre-trained networks is to avoid overfitting. Kahou et al. [16] discussed the advantages of pre-trained models. The multi-stage fine-tuning method can further boost the performance of the FER. Multiple feature input networks are designed to tackle the problems of image rotation, scaling, and illumination effects. Instead of feeding normal RGB images, handcrafted features like Scale-invariant feature transform (SIFT) and mapped local binary pattern features are given as input to the deep neural networks.

Spare blocks and layers based deep neural networks are used to improve the performance of FER. A novel loss function known as the center loss is introduced to improve the discriminative power of CNN. Center loss [17], along with the softmax layer, is used at the end of the CNN layer to obtain a good threshold for classification. Many loss functions like island loss [18], and triplet loss [19] are deployed into CNN models to boost the discriminative power of FER. Ensemble-based deep neural networks can be again classified into multi-architecture ensembles, feature level ensembles, and decision-based

ensembles. Multi-architecture ensemble models [20] use the different error functions like log-likelihood loss and hinge loss to feed weights to respective networks inside the model adaptively. Feature level ensembles concatenate important features derived from different networks in the model into a one-dimensional feature matrix. Decision-based ensembles adapt classification based on rules like majority voting [21], simple average [21], and weighted average [20]. Generative adversarial networks (GANs) are widely used in recent years in the field of facial expression recognition. The models trained with GANs can perform image synthesis, which is realistic and accurate. They can overcome class imbalance issues in different datasets by adding more training images to the dataset. Zhang et al. [22] introduced a GAN-based FER model to synthesize images with various expressions under random poses for multi-view facial expression recognition.

The use of deep learning techniques has increased in various computer vision problems in recent years. DNN models such as convolutional neural networks and recurrent neural networks are commonly used in pattern recognition tasks. In this section, we briefly discuss existing hybrid models on FER.

Multi-feature fusion models:

Tang et al. [23] proposed a two-feature fusion approach. The first model, the differential geometric fusion network, derives handcrafted features based on Euclidean distances between important facial landmarks. The second model, known as the multi-dimensional convolutional neural network, extracts the deep features. The effectiveness of the combination of two features was evaluated on the CK+ dataset for six emotion classes (anger, surprise, disgust, happy, fear, and sad).

Wang et al. [24] proposed a weighted fusion of two hand-engineered features. Multi-scale block local binary pattern uniform histogram and HOG features are initially derived from the images and fused. The fusion model uses an SVM to classify emotions. The model classifies seven emotions (angry, happy, sad, surprise, contempt, fear, and disgust). The paper also explains the variations of unweighted and weighted fusion of features.

Xiaohua et al. [25] used Weber Local Descriptor (WLD) and HOG to create a hybrid model for facial expression recognition. The experiment explains that the combination of Weber Local Descriptor and HOG provided better results when compared with other

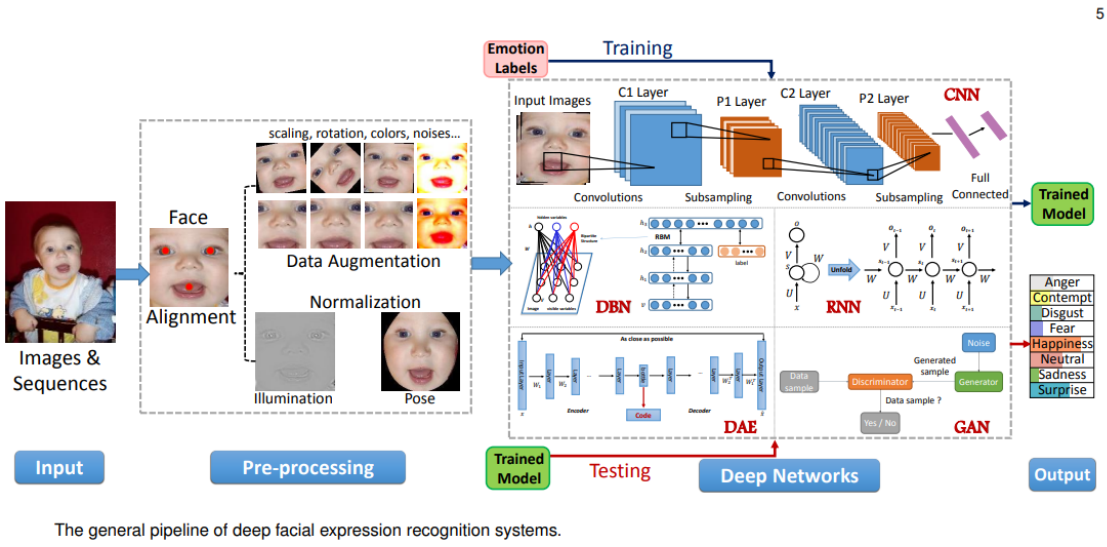

Figure 2.2 General Pipeline of FER models [2]

popular features like Gabor Wavelet, LBP (local binary pattern), and AAM (Active Appearance Model). WLD extracts the edges and texture detail and is very consistent with human perception. The studies show that the model is unaffected by noise and non-monotonic lighting differences. Xie [26] has proposed a FER model that combines spatially maximum occurrence model (SMOM) and elastic shape–texture matching (ESTM). The SMOM is based on the statistical characteristics of training facial images, and ESTM measures the similarity between images based on texture and shape regions. The fusion of two features has significantly improved the classification accuracy. Lin et al. [27] have proposed a multi-feature fusion of two-dimensional principal component analysis and local binary pattern (LBP). The paper explains that the local texture is represented by a local binary pattern (LBP) to extract global appearance features. The two inputs are then given to the acyclic graph (DDAG) based support vector machine (SVM) for identifying several prototypic facial expressions.

Viswanatha et al. [28] proposed a similar hybrid model that comprises deep and handcrafted features. The XceptionNet model is used to extract the deep features, and for the handcrafted features, the authors have used facial landmarks using OpenCV. Pan et al. [29] proposed fusing HOG and CNN spatial-temporal features for video-based FER. The authors have used VGG-16 architecture and HOG to extract temporal and spatial characteristics. The authors discuss the importance of fusing CNN and HOG features to extract local and abstract features of facial regions.

The proposed work in this research also uses a multi-feature fusion model but differs in architecture, selection of features, method of concatenation, and hyper-parameters selection. The existing works have used different feature fusion methodologies, for example, Xiaohua et al. used two hand-engineered feature fusion (WLD+HOG), Wang et al. used LBP+HOG in the FER design.

Many existing models use unique techniques in detecting facial expressions. Breuer and Kimmel [30] have evaluated different CNN visualization strategies and addressed the potential of DNNs to understand emotions. Jung et al. [31] have used two different CNN models to improve the FER accuracy. The first CNN extracts temporal information from their respective image sequences, and the second CNN considers facial landmarks to obtain information regarding temporal geometry. The definition of the deep region was introduced by Zhao et al. [32] through multi-label learning (DRML), which utilizes feed-forward networks to understand facial regions and evaluate the structural patterns of the face by forcing the knowledge to be captured by learned weights. In the FER method, neural networks that use pre-trained networks have been implemented, and models have been deployed to minimize training time. The aim of using these pre-trained networks is to use pre-trained weights trained on large datasets such as Imagenet [33]. The benefits of choosing pre-trained networks are identified by Kahou et al. [34] in the proposed work. Koc [35] has proposed the Sum of Pixel Slope Similarities approach for the face recognition task. The authors have compared the approach with other popular recognition models like the common vector approach (CVA), discriminative CVA, and support vector machines. The work concludes that the recognition rate is improved when using the surface normal vectors rather than the gradient vectors in each pixel. Liu et al. [36] proposed a new Boosted Deep Belief Network (BDBN) for iteratively executing the three training stages in a system. A selection of features that effectively characterize expression-related facial appearance/shape changes can be learned and selected using the proposed BDBN system. Liu et al. [37] also proposed the idea that deep belief networks are also used to train the face parsing detectors, which are then fine-tuned using logistic regression. Mollahosseini et al. [38] proposed a network that has two convolutional layers, each with max-pooling and four Inception layers in between. The network is a single-component architecture that takes registered facial images as input and categorizes them into six basic or neutral expressions. Korrami et al. [39] presented a method for determining which

parts of the face affect CNN's predictions. To analyze the facial expression information of temporal sequences, a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) was proposed by Xhang et al. [40]. Kurup et al. [41] proposed an algorithm that employs a cascaded structure in which the facial images are first subjected to feature extraction, accompanied by feature reduction. A Deep Belief Network (DBN) is trained semi-supervised using all available labeled and unlabeled data, and using a reconstruction error-based rating, features are selected to exclude those that do not provide details. Datta et al. [42] proposed a model in which concatenation of geometric and texture-based features is used to create a fast facial emotion classification system. Cai et al. [43] proposes a novel island loss to enhance the discriminative power of deeply learned features in a FER.

Kim et al. [44] discussed FER based on ensemble methods. The ensemble techniques like feature-based ensembles and decision-based ensembles are also discussed in the proposed work. To adaptively update the weights of neural networks, these multi-architecture ensemble models use distinct cost functions such as loglikelihood loss and hinge loss. Feature level ensembles fuse and build critical features taken from different networks into a one-dimensional matrix of features. Other classification methods, such as majority voting [45], simple average, and weighted average [46], are also used in designing a FER model. The advanced FER methods use generative adversarial networks (GANs) that can create synthetic realistic machine-made patterns. In recent years, GAN-based frameworks were designed to develop machine-made synthetic images for multi-view facial expression recognition with varying illumination for random poses.

FER has been investigated in the computer vision field for decades [47], [48]. According to the existing FER models, the approaches may be divided into two categories that are Static image based and Dynamic image sequence-based approaches. In recent years, the usage of deep learning approaches in different computer vision challenges has risen. In pattern recognition tasks, Deep Neural network (DNN) models such as convolutional neural networks and recurrent neural networks are often utilized. In this part, we will go through some of the existing FER models.

Zhang et al. [49] suggested one of the still image-based approaches on the FER-

2013 Challenge [50], the authors utilized a deep CNN accompanied by a linear one-vs-all support vector machine (SVM) and obtained good classification accuracy. Yu et al. [51] suggested an emotion detection module based on an ensemble of several networks, each with a separate set of weights. Breuer and Kimmel [52] investigated the potential of DNNs to grasp emotions by evaluating several CNN visualisation methodologies. Jung et al. [53] improved FER accuracy by employing two separate CNN models. Zhao et al. [10] defined the deep region using multi-label learning (DRML), which uses feed-forward networks to understand facial regions and assess structural patterns of the face by forcing knowledge to be captured by learnt weights. Mollahosseini et al. [54] suggested a network with two convolutional layers, each with max-pooling, and four Inception layers between them. The network is a one-component architecture that takes in captured facial images and categorises them into six basic expressions along with neutral expression. The FER approach employs neural networks that leverage pre-trained networks, and models are deployed to save training time. The goal of using these pre-trained networks is to employ weights that have been developed during training on huge datasets such as Imagenet [33]. Kahou et al. [55] revealed the advantages of using pre-trained networks. It is worth noting that the temporal relationship between image frames in sequence is critical for detecting face emotions. Recently, there has been a greater emphasis on methods that capture spatial–temporal aspects [[56], [57], [58]]. For video-based expression recognition, Liu et al. [59] employed a 3D-CNN architecture. They suggested a CNN architecture with flexible facial action parts model constraints that can learn spatial–temporal properties as well as locate facial action parts. For FER, Khorrami et al. [60] built a CNN–RNN architecture. They also looked at how much each network adds to the framework. Jaiswal et al. [58] proposed a model for obtaining temporal information using a mixture of CNN and BiLSTM, which outperformed other models in terms of accuracy. Fan et al. [57] developed a hybrid network that extracted features using a 3DCNN architecture and then utilized RNN to capture the temporal relationships for FER. According to the preceding discussion, multiple network integration and CNN–RNN frameworks considerably increase FER performance. The objective of the proposed work is to learn discriminative spatial–temporal features, particularly temporal motion context information using VGG-Face CNN-LSTM based architecture.

Chapter 3

Facial Expression Recognition using a Dual CNN Model with Novel LogicMax Layer

3.1 Introduction

The existing FER models fail to classify expressions on micro-feature level. The micro-expressions play an important role in classifying emotions. The high intraclass correlation between expressions makes the classification task harder. Facial expressions convey important features for recognizing human emotions. It is a challenging task to classify accurate facial expressions due to high intra-class correlation. Conventional methods depend on the classification of handcrafted features like scale-invariant feature transform and local binary patterns to predict the emotion. In recent years, deep learning techniques are used to boost the accuracy of FER models. Although it has improved the accuracy in standard datasets, FER models have to consider problems like face occlusion and intra-class variance. In this research work, we have used two convolutional neural networks which use vgg16 architecture as a base network using transfer learning. This chapter explains the method to tackle issues of classifying high intra-class correlated facial expressions through an in-depth investigation of the Facial Action Coding System (FACS) action units. We have used a novel LogicMax layer at the end of the model to boost the accuracy of the FER model. Classification metrics like Accuracy, Precision, Recall, and F1 score are calculated for evaluating the model performance on CK+ and JAFFE datasets. The model is tested using 10-fold cross-validation and obtained classification

accuracy rate of 98.62% and 94.86% on CK+ and JAFFE datasets respectively.

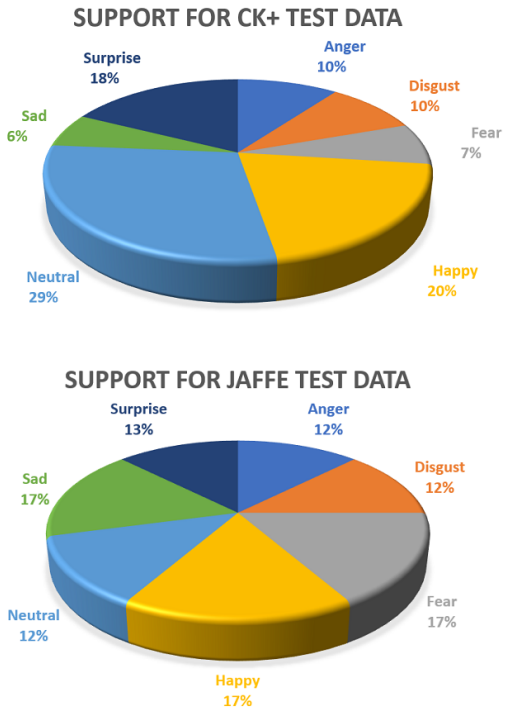

Figure 3.1 Information about the samples used in test data for different classes

3.2 Role of FACS in the Proposed CNN

Deep Neural Networks like Convolutional Neural Networks (CNN) try to derive both low and high-level features automatically through training with datasets. Low-level features are important lines, edges, and corner points, which can be extremely useful in predicting the overall class. The initial stages of a CNN extract the low-level features, and as we move deeper into the network, it tries to combine these low-level features into a meaningful class. Facial Emotions are tough to classify since the problem is a sub-classification task, which involves identifying the emotional classes that have a very slight variance. Paul Ekman and Wallace V. Friesen developed a system known as Facial Action Coding System (FACS) [7] that identifies different facial expressions on any human face. The important facial features are deconstructed and properly taxonomized according to their property using FACS. FACS helped to generate and classify independent actions of muscles/muscle contraction and relaxation known as “Action Units” (AUs). A combination of different action units on the face denotes a particular emotion, as shown in

Table 3.1 FACS action units for different emotions. [7]

Emotion	Facial Muscle	Corresponding Action Units
Anger	Brow lowerer+ Upper lid raiser+ Lid tightener+ Lip tightener	4+5+7+23
Disgust	Nose wrinkle+ Lip corner depressor+ Lower lip depressor	9+15+16
Fear	Inner brow raiser+ Outer brow raiser+ Brow lowerer+ Upper lid raiser+ Lid tightener+ Lip stretcher+ Jaw drop	1+2+4+5+7+20+26
Happiness	Cheek raiser+Lip corner puller	6+12
Sadness	Inner brow raiser+Brow lowerer+Lip Corner depressor	1+4+15
Surprise	Inner brow raiser+Outer brow raiser+ Upper lid raiser(Slight)+ Jaw drop	1+2+5B+26

Table 3.2 Intensity level classification of action units in FACS. [7]

Alphabet	A	B	C	D	E
Intensity Level	Trace	Slight	Pronounced	Extreme	Maximum

Table 3.1. Each emotion triggers different facial expressions, and if the FER model tries to analyze the facial expressions accurately, then the classification of emotions becomes easy. Table 3.1. explains the importance of different action units for corresponding emotions.

FACS has scaled the intensity of the action units by introducing levels from A to E, where A is the weakest and E as the strongest intensity, as shown in Table 3.2. From the Table 3.1., it is evident that various emotional states have the same facial muscle moments, for example, Disgust and Sadness emotions trigger the Lip Corner Depressor (Action unit- 16). There is a reasonable probability of misclassifying the emotions due to these similarities in the different emotion classes. FACS helps in modelling an excellent deep neural network by exposing the correlation between emotions. There is much difference in the emotion class happiness and surprise because there is no intersection of action units in both the classes. FACS can convey important information regarding the probability of differentiating two emotions through the study of their respective action units. We have designed a new architecture that uses FACS information along with dual CNN in predicting the emotion class. The inclusion of FACS information in the CNN model improved the accuracy of the model and helped in a better understanding of the role of action units in emotion classification. On analyzing the action units in emotions, the majority of them lie on the crucial facial landmarks like eyebrows AU (1,2,4), eyes AU

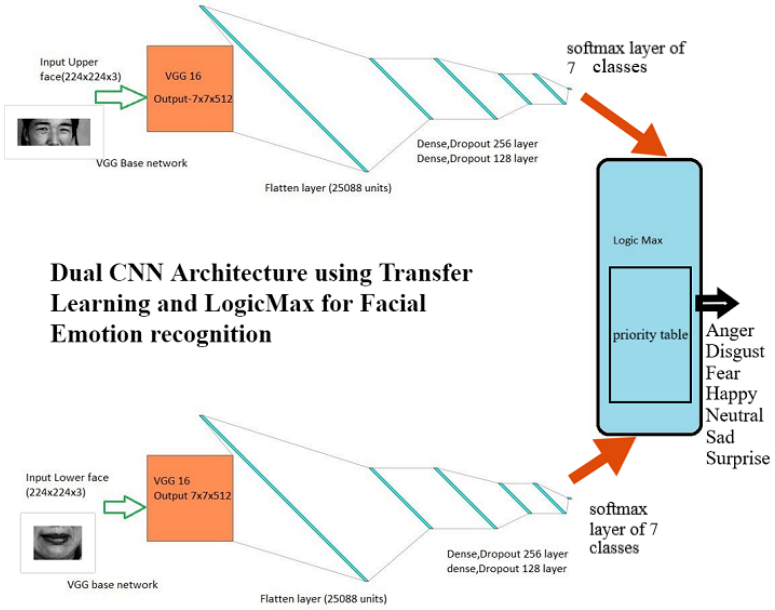

Figure 3.2 Overview of the proposed architecture

(5,7), and lips AU (23,16,26,12,23). A face can be symmetrically divided into two parts, either vertically or horizontally. When a face is split vertically, two images share identical action units since both are mirror images. When a face is horizontally split, we obtain two asymmetric images that have different landmarks. The upper half has eyebrows, eyes, and nose as essential landmarks, and the lower part of the face has a mouth and chin as crucial landmarks. To improve accuracy, we designed two separate deep convolutional networks to identify the emotion on both the upper and lower parts of the face. Each convolutional neural network tries to extract different features in their respective sections (upper or lower region of the face). In doing so, CNN models can spatially concentrate on feature extraction of their respective landmarks. This complex architecture improves the efficiency of the model and also traces each landmark's behavior in different emotions. Some action units are more pronounced when compared to others, for example, emotions like surprise. In some situations, the emotion predicted on the upper face contradicts the emotion predicted on the other half of the face. During the mismatch, there has to be a logical conclusion on the emotion of the subject. A new layer called "LogicMax" helps to solve the problems of mismatch by building a priority table. LogicMax is a layer that is fitted at the end of the two CNN models to take a logical conclusion of the emotion class of the overall face. LogicMax is a novel layer that predicts the final emotion class of the overall face by analyzing both the outputs of two CNN. Thus, FACS information

along with a logical approach can be imparted inside the logicMax layer to improve the efficiency of classification.

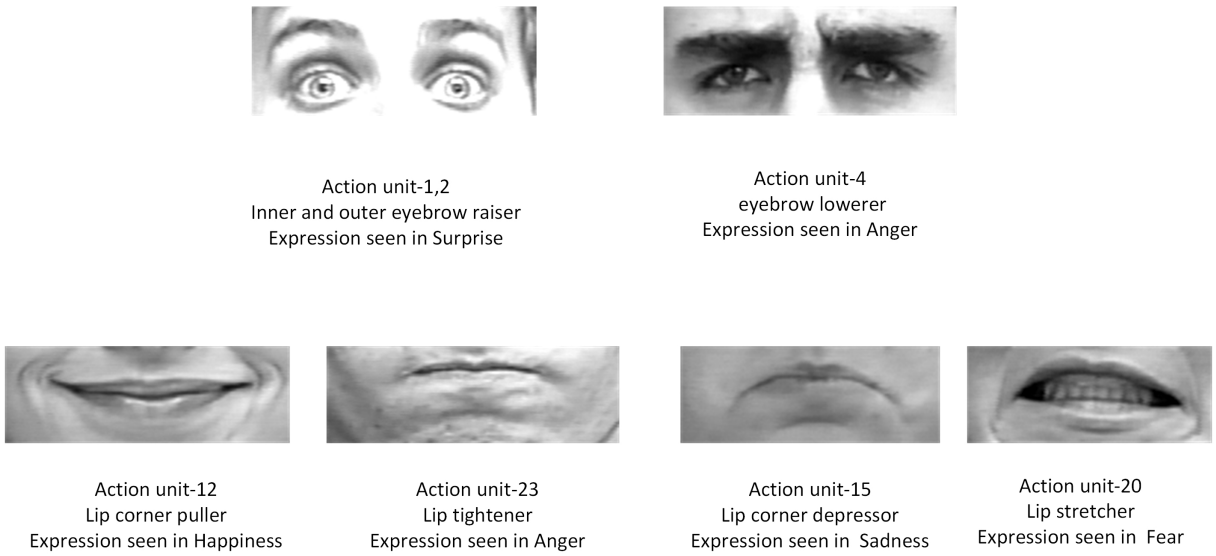

Figure 3.3 Action units observed in few facial expressions

3.3 Design of the Proposed FER Model using Transfer Learning and Logic-Max

ImageNet [61] is one of the knowledge transfer projects which provides huge datasets that are useful for training models. Powerful models like Inception, VGG-16, and Resnet are trained on ImageNet data, which consists of thousands of image categories. As the models are pre-trained with a huge database, they have a good ability to extract the features like edges, corners, and different shapes. It is wise to implement these models on our problem statement as it saves a lot of computation and time.

The proposed model, as shown in Fig. 3.4 has two CNN architectures that use VGG 16 architecture [13] as its base network through transfer learning. The pre-trained network gives better feature extraction and also saves much time when compared to training a whole new model. The pre-trained weights of the VGG16 network are loaded into their corresponding convolutional filters. There are many advanced pre-trained models like Resnet 101, Inception V2, and Inception V3, but VGG 16 is selected in this work as it has a good trade-off between loading time vs. feature extraction [62]. The proposed work

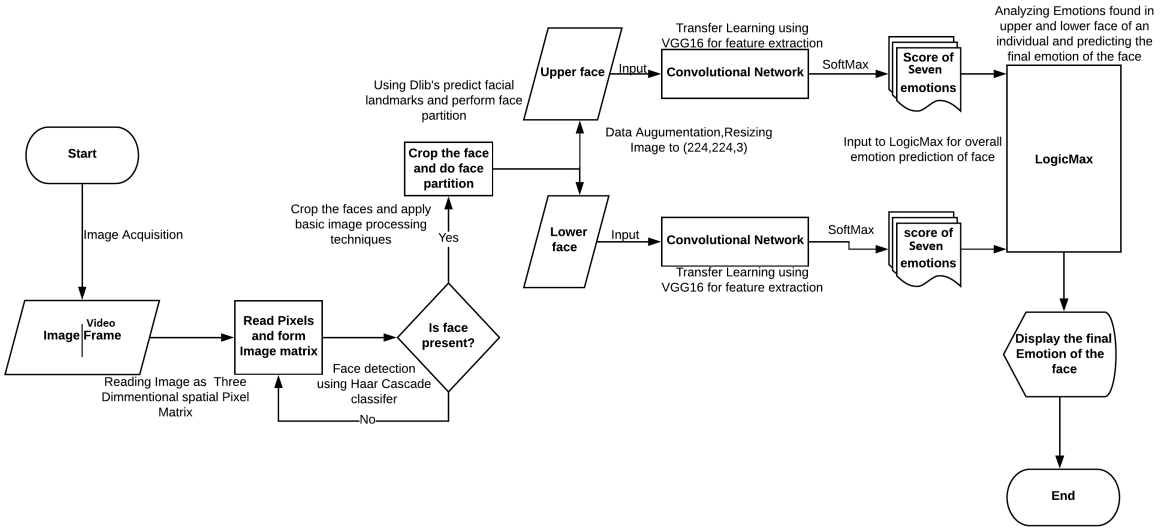

Figure 3.4 Flow chart of the model process

considers dual CNN architecture so, the VGG 16 has been chosen for its simplicity and fastness. The VGG16 base-network weights are disabled from training since it has good pre-trained weights for feature extraction.

To the base network, we have added a flatten layer, a dense layer of 256 neurons, and a dropout layer. It is followed by another dense layer with 128 neurons, a dropout layer, and a softmax layer of seven output classes, refer to the model design in Fig. 3.5. The training of the network involves only updating the weights of the added layers to the base network. This same model is used twice to determine the emotion class on the upper face and lower face. We have implemented this model using the Keras framework. In the Fig. 3.2. representation of the model with the VGG16 as a base network and other added layers is shown.

3.3.1 Partition of Face

The research work considers lower and upper face parts for emotion prediction. The research in the study of facial emotion analysis in humans has revealed that the eye and mouth movements alone play an important role in the display of micro expressions [63]. The upper half of the face considers the eyes as an essential indicator of emotion analysis and the lower half considers the mouth as an important landmark to understand facial

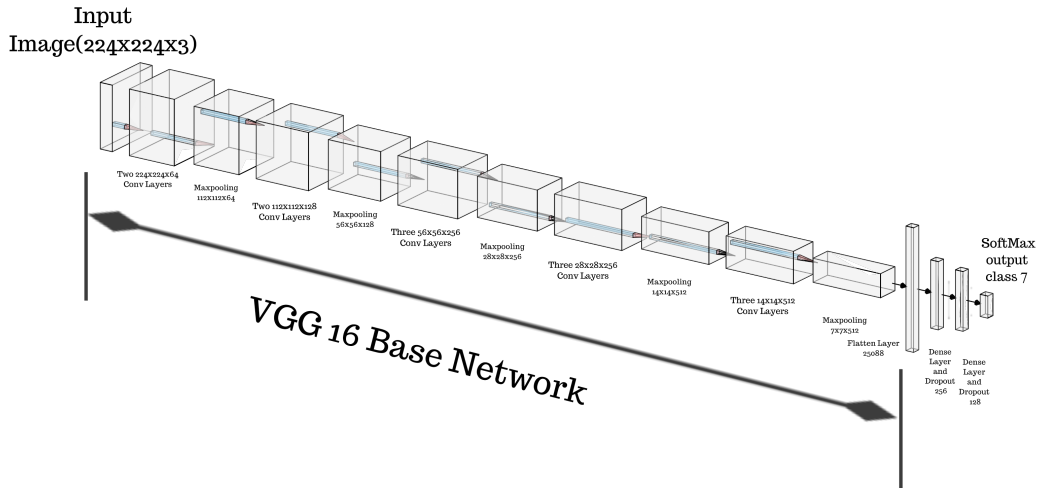

Figure 3.5 Structure of single convolutional neural network used in the model

Figure 3.6 Filter visualization from both CNN models

expression. The face can also be divided into 3 or 4 parts but the increase in the division increases the number of CNN models which can complicate the practical design of the FER model. Localization of faces on image or video frames is done using Haar Cascade Classifier. The extracted face is to be partitioned into the lower and upper parts. The partition of the face into more than two parts can make the algorithm complex and highly computational. We have used the Dlib library [64] and Open CV tools [65] to partition the face. Dlib's facial landmark detector is a helpful tool to identify important facial landmarks like eyes, nose, eyebrows, and jawline. Facial landmark extraction using Haar cascades is also possible but, the training to detect landmarks requires huge training of the classifier with positive and negative images to produce an accurate cascade classifier for landmarks. Kazemi and Sullivan [66] have implemented a facial landmark detector using an ensemble of regression trees. The algorithm can produce 68 coordinates of important

facial landmarks. The coordinates of point 33 lie on the exact center of a face, as shown in 3.7. We have successfully implemented a program using Opencv tools to partition a face using the dlib landmark information. The pixels that lie above point 33 belong to the upper part of the face, which includes the eyes, nose, and eyebrows pixels, and the pixels that lie below point 33 belong to the lower part of the face, which includes the mouth as an important landmark. The two face parts are given as inputs to their corresponding CNN models for determining the emotion class, refer to Fig. 3.2.

Figure 3.7 Partition of face

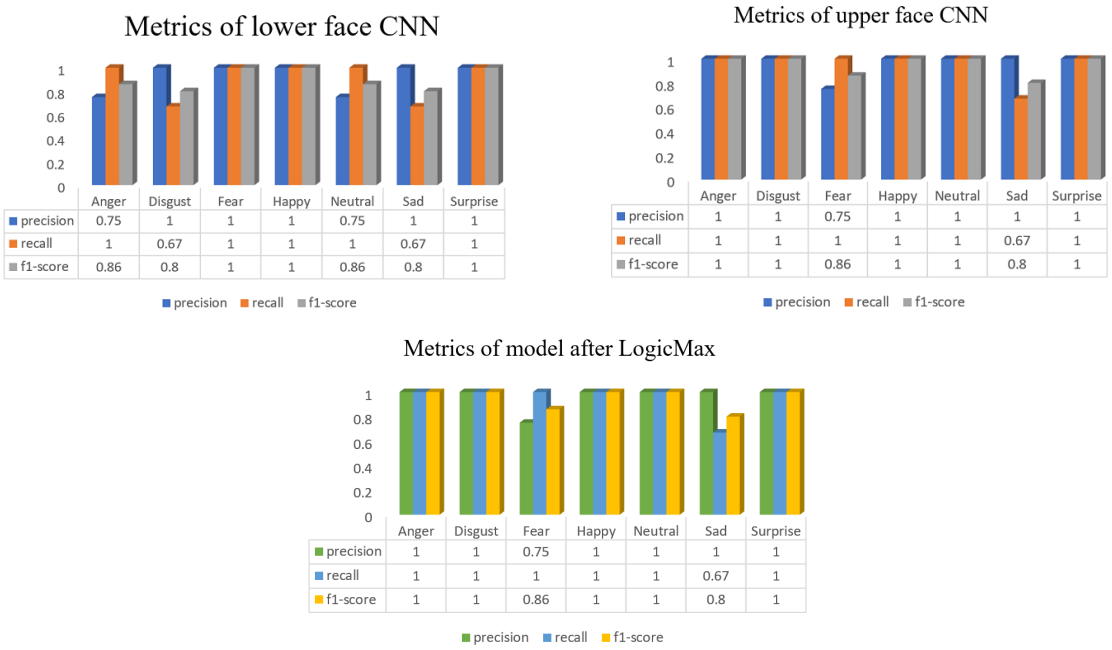

Figure 3.8 Classification metrics of the model on JAFFE test data

3.3.2 Data Augmentation

The extracted face parts are data augmented with unique parameters. Data Augmentation is a technique used to create artificial images from the dataset by transforming

Table 3.3 Information about the k-fold validation process on CK+ and JAFFE datasets.

Dataset	Total Samples	Training Samples	Testing Samples	Method of Testing	Data Selection
CK+	1479	1331	148	10 -fold cross validation	last three to four samples of a sequence (includes peak expression frame)
JAFFE	213	192	21	10 -fold cross validation	All the samples from the dataset are taken for training and testing

the image geometrics and adding random noise. The important image transformations done in data augmentation are Rescaling, Rotation, Shear, Zooming, Width Shifting, Height Shifting, Horizontal flipping, and Vertical Shifting. The combination of different parameters is to be carefully chosen to generate a good synthetic dataset. Data Augmentation is useful to eliminate the overfitting problem [67]. Overfitting in machine learning occurs when the model tries to memorize the patterns instead of learning to detect complex patterns in the training data. Detection of facial expressions should be robust even in case the image is tilted, mirrored, or zoomed. Data augmentation should be carefully performed since it can also lead to serious underfitting problems. Generally, the training and validation error helps in analyzing overfitting and underfitting problems in deep neural networks. If the model has good training accuracy but has very less validation accuracy, then the model is overfitting to the data. If the training accuracy is very less than that of validation accuracy, then the model is undergoing underfitting. It is seen that the width shifting of the training set during data augmentation does decrease the accuracy of the model since the important landmarks get affected by high width shifting so, width shifting is not performed during the data augmentation. Only the training set is data augmented; the validation dataset is not data augmented but only rescaled. Table 3.4. shows the magnitude of variations of each operation during the process of data augmentation. Various combinations of values are applied, and the data are shown in the Table 3.4. gave us the best results, and overfitting problems are avoided through the proper data augmentation process.

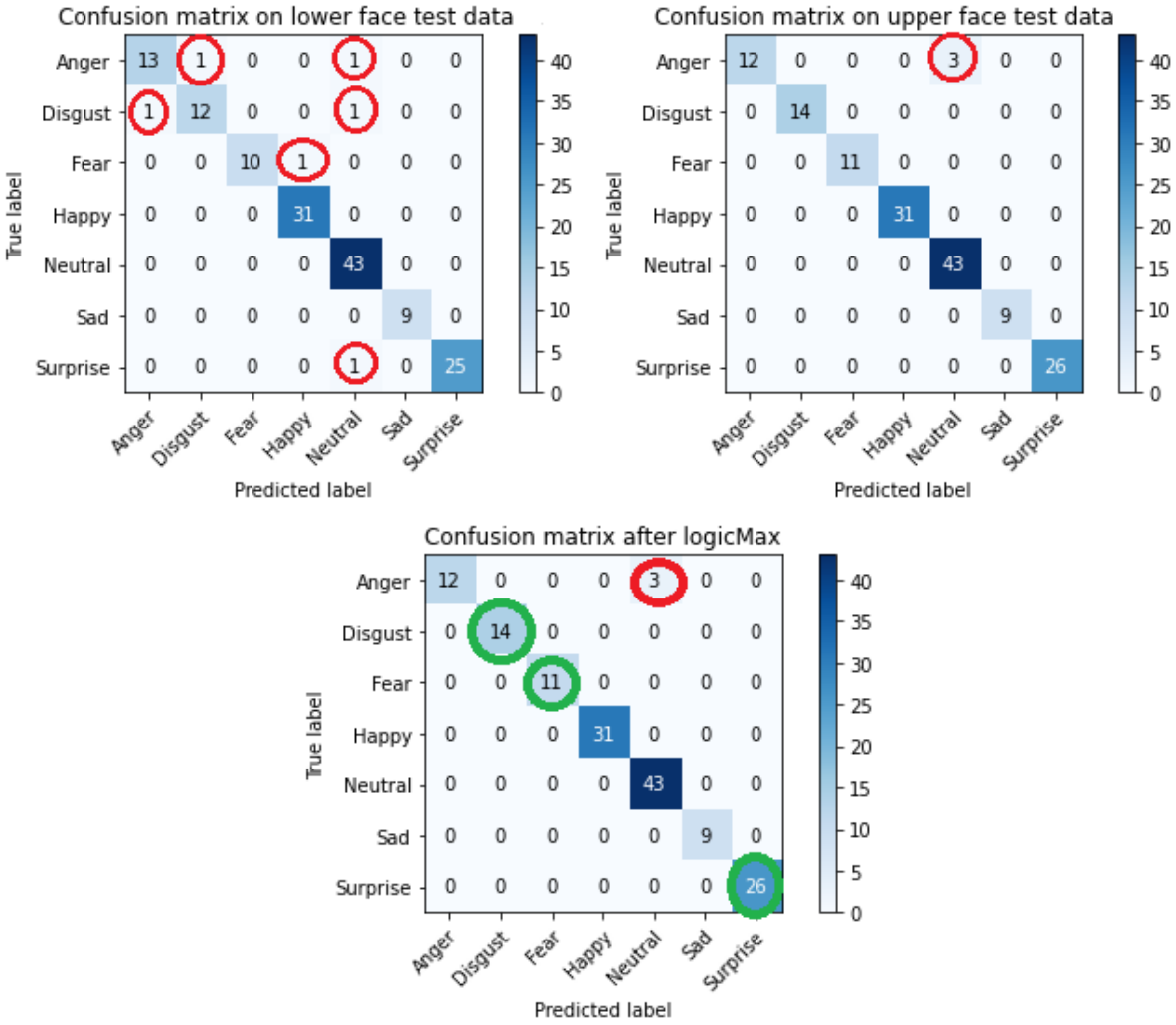

Figure 3.9 Confusion matrices on third fold CK+ test data during 10-fold cross-validation

3.3.3 Training and Validating the Model

We have used Google's Collabalortoy GPU to train our model. Google's Colab provides GPU Nvidia 1xTesla K80, having 2496 CUDA cores and CPU Xeon Processor of the frequency of 2.3 GHz. Input images are resized to (224x224) as the VGG16 model is trained for (224x224) sized images. The RMSprop optimizer is used in training the model. The loss function categorical cross-entropy is used as an error function for training the weights of the neural layers. The cross-entropy loss function is widely used in classification problems for deep neural networks [68]. For each batch input of images, the softmax layer produces the predicted outputs which contain CNN scores of all emotion classes. The softmax layer is a function that transforms arbitrary random values into a properly ordered probability distribution. SoftMax layer function gives output ranging between

Table 3.4 Data Augmentation on training set.

Operation	**Scaling factor**
Rescale	1/255
Rotation Range	0 - 30
Shear Range	0 - 0.15
Zoom Range	0 - 0.15
Height shift Range	0 - 0.2
Width shift Range	No
Horizontal Flip	Yes
Fill mode	Nearest

(0,1). The total number of classes present in the CNN model is 7. Let us consider t_i and y_i be the target and the softmax score of i^{th} class of a sample.

$$Softmax\ score\ for\ each\ class\ i=\ 1\ to\ 7 : f(y)_i = \frac{e^{y_i}}{\sum_j^{N=7} e^{y_j}} \tag{3.1}$$

$$Categorical\ Cross\ entropy\ error : -\sum_{i=1}^{N=7} t_i log(y_i) \tag{3.2}$$

The sum of all outputs from the softmax layer equals one. In Multi-Class classification problems, the targets are one-hot encoded, which makes only the positive emotion class appear in the categorical loss function.

3.3.4 Exploratory Data Analysis of FACS Action Units and Emotions

The dual CNN architecture predicts the emotion on two respective inputs (upper and lower faces). The next problem after the classification is the mismatch between the emotions. If the prediction of two trained CNN models mismatch it defines that the expressions shown in the two spatial regions show discordance in emotions. This discordance shows the complexity of the facial expression. This problem is solved by the LogicMax layer. The logicMax layer carefully analyzes the two outputs and finally concludes with a single emotion on the overall face. The priority table built inside the logicMax layer de-

Table 3.5 One hot encoding of action units for six emotions.

Action Units	**Angry**	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**
AU1	0	0	1	0	1	1
AU2	0	0	1	0	0	1
AU4	1	0	1	0	1	0
AU5	1	0	1	0	0	1
AU6	0	0	0	1	0	0
AU7	0	0	1	0	0	0
AU9	0	1	0	0	0	0
AU12	0	0	0	1	0	0
AU15	0	1	0	0	1	0
AU16	0	1	0	0	0	0
AU20	0	0	1	0	0	0
AU23	1	0	0	0	0	0
AU26	0	0	1	0	0	1
AU27	1	0	0	0	0	0

cides the final emotion based on certain conditions. The following information is essential in designing the priority table.

The LogicMax is an important layer added to the end of two CNN models. The CNN models predict the emotion class of their respective inputs (lower and upper face). The action units shown in Table 4 are sufficient for analyzing the emotions since the FACS considers these action units important for predicting the emotions on the face, refer to the Table 1. If both the lower and upper face CNN models predict the same emotion class, there is no perplexity involved in the decision-making of the overall emotion of the face. But, if the lower and upper face CNN models predict different emotion class, there has to be a logical conclusion on the overall emotion of the face. This logical conclusion can be sought by doing exploratory data analysis on the different action units of emotions. The correlation between different emotions on the lower and upper face provides an important basis for designing the logicMax layer. The spatial distribution of important action units of emotions is shown in Fig. 3.10. The combination of the action units is shown in the Table 3.5. form important basics in identifying the different emotions according to FACS (refer to the Table 3.1). The categorical action units are one hot encoded for correlation

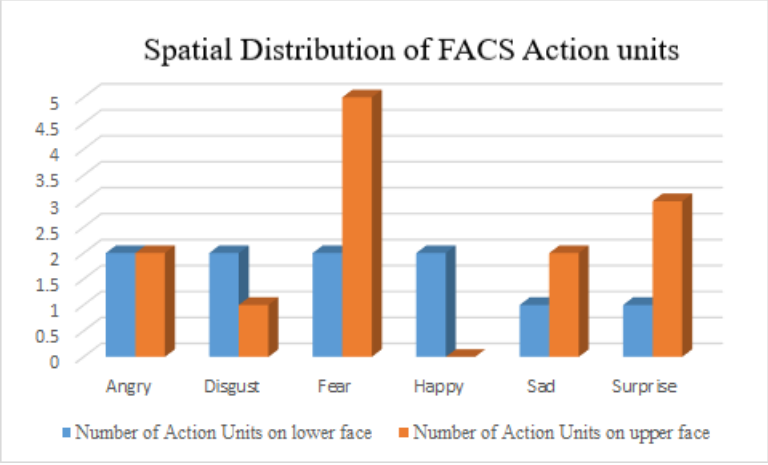

Figure 3.10 Spatial distribution of action units on face

analysis which is shown in the Table 3.5.

There are no action units present on the upper face for emotion happiness. The emotion of happiness has all the crucial facial action units present on the lower face (cheeks and mouth). Hence, according to FACS, if the lower CNN model predicts happiness, then it is not required to investigate the emotion class of the upper face. The correlation of

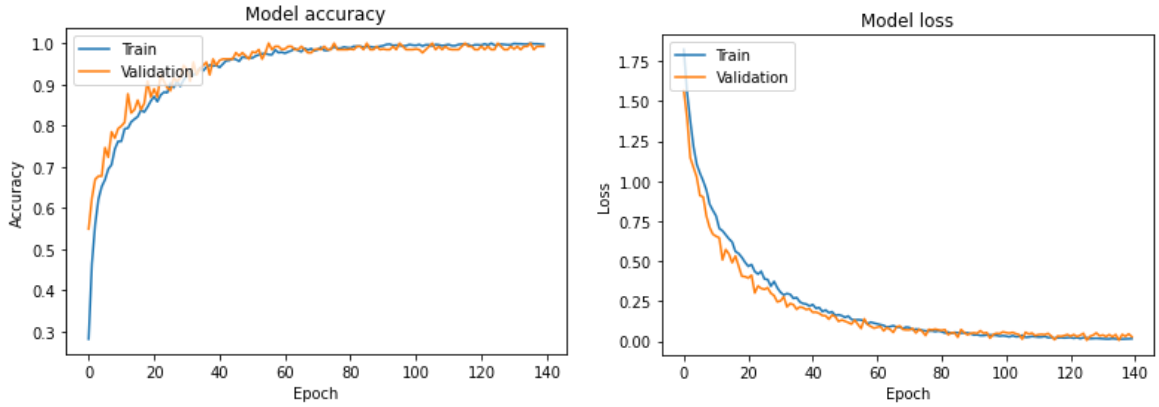

Figure 3.11 Train and Validation graphs of upper face CNN model on CK+ dataset during 10-fold cross-validation process

emotions in the lower and upper face is analyzed using the Pearson correlation matrix. It is clear that on the lower face, the emotions like Disgust and Sad have a good correlation since they share the same action unit 15 (Lip corner depressor) refer to Table 3.5. Therefore, there is a high probability that emotion disgust can be predicted as sad and vice versa by the lower face CNN model. In this case, it is essential to observe the upper face CNN model's prediction. It is clear that on the upper face, the disgust emotion has a unique action unit 9 (Nose wrinkle), refer to Table 3.1. We can also observe the disgust emotion has a poor correlation with all other emotions on the upper face, as shown in the correlation matrix refer Fig. 3.12 So, in the case of a mismatch, if the upper face predicts a disgust emotion class, there is no need to investigate the emotion of the lower face. We have designed priorities in predicting the final emotion in case there occurs a mismatch between the prediction of two CNN models, refer to Table 3.6. The value of the correlation coefficient lies between -1.0 and 1.0. The value of the correlation coefficient determines the power of association. If the value of the correlation coefficient lies between 0.5 and 1.0, it suggests a strong positive association. The correlation coefficient between 0 and 0.5 suggests a weak positive association. The correlation coefficient below 0 to -1.0 indicates a negative correlation.

3.3.5 LogicMax And Priority Table

The LogicMax layer is a novel decision-making layer introduced in this model. The logicMax, unlike softmax, can be tuned and modified according to the nature of the

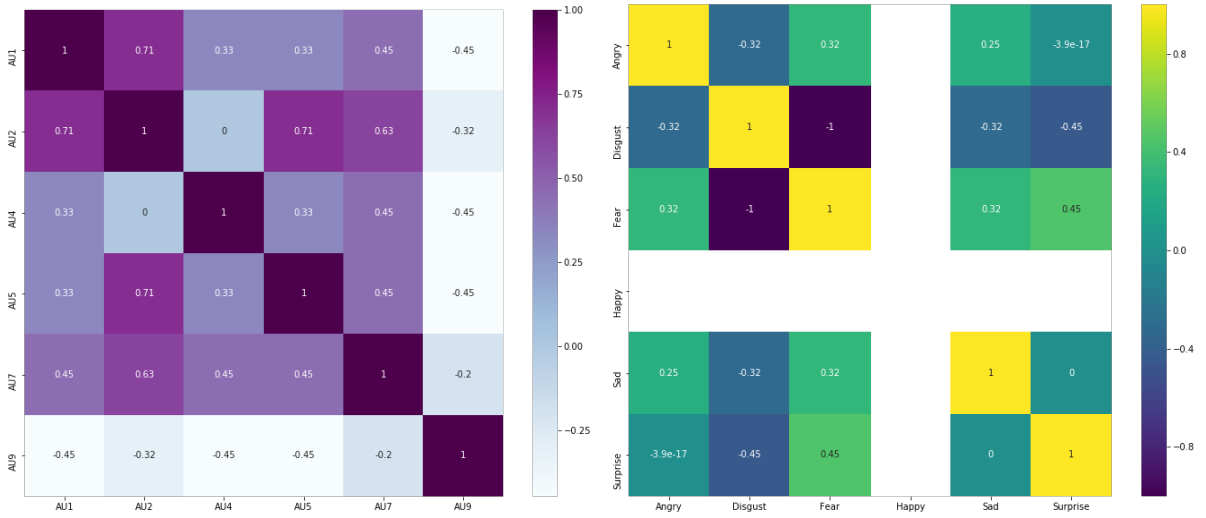

Figure 3.12 Correlation of different action units and emotions on upper face

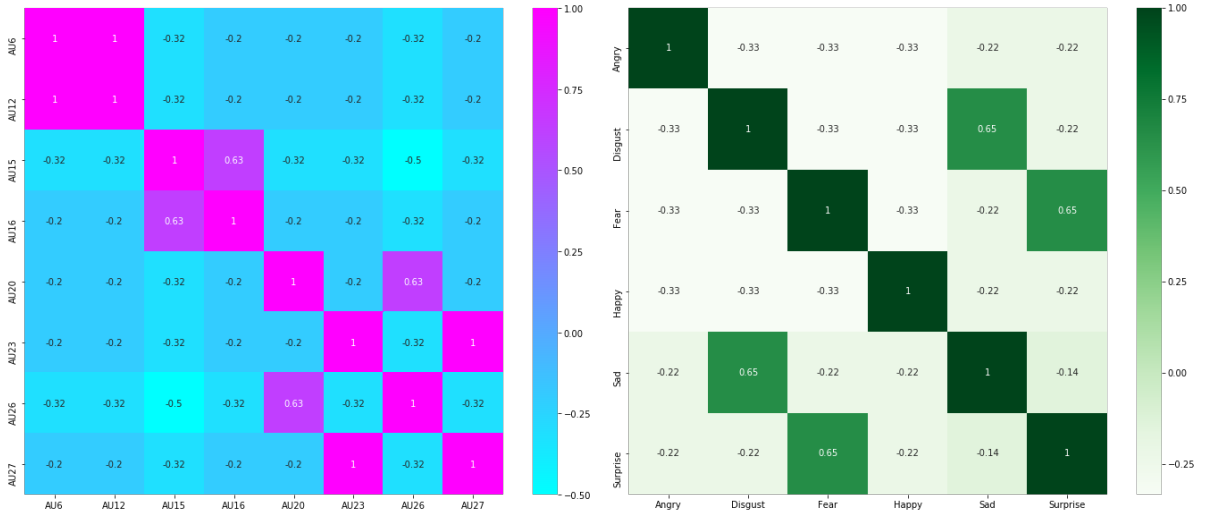

Figure 3.13 Correlation of different action units and emotions on lower face

output class. The logicMax aims to impart human intelligence and logical thinking inside a CNN model. In the proposed model, the function of logicMax is to predict the overall emotion seen on a face by analyzing the emotions found in the lower and upper face regions. The real discriminative power of logicMax is utilized in the situation when the emotions predicted by the two CNN models mismatch. The mismatch is often seen in facial expression classification tasks since the correlation is high among different emotions. In these situations, the layer should choose any one of the two CNN model outputs. This prioritizing should be thoroughly performed by analyzing different features that appear in facial expressions. A set of rules is framed inside the priority table that decides the output class by analyzing the features. For creating a priority table, three types of features are

considered. The three types of features are explained in the below points.

- **Type-1 features**: Analyze and locate the unique features present in different emotions. In the one-hot encoding Table 3.5., emotions like happiness and disgust have unique action units (12,9), respectively. Action unit 12 in happiness is found on the lower face, and action unit 9 in disgust emotion is found on the upper face. These action units are unique since they are not found in other emotions. These features are given the highest priority in the LogicMax. When a mismatch of emotion class between the two CNN models at the softmax layer occurs, these unique features are examined primarily in the input. These features can be termed "Type-1 features". The correlation heat map charts shown in Fig. 3.13 and one hot encoding Table 3.5. provides important information about the type-1 features.

- **Type-2 features**: The features which have a poor correlation with all other features are the next vital patterns that need to be examined in the LogicMax layer. If the unique features (type-1) are not available in the input, these types of features are explored in the input. These features can be easily discriminated against as they have a weak correlation with other standard features. The anger class has an action unit 23 present on the lower face, which has a weak correlation with other action units, refer to Fig. 3.13. The detection of anger class in the lower face suggests there is a higher probability that the subject displays anger emotion. These features can be termed "Type-2 features".

- **Type-3 features**: The next type of features are not so important as the type-1 and type-2 features since they are trivially found among different emotions. These features are given less importance in the priority table as they are not unique and appear in two or more emotions. For example, action units 4 and 5 can be observed in emotions like anger, fear, surprise, and sadness. These features create perplexity to the classifier as they are found in different classes and have a chance to increase the correlation among different emotion classes. These features can be termed "Type-3 features". If any face expresses these unique features then the priority table gives high preference to these features. If the priority table finds disgust in the upper face and sad in the lower face, the priority table decides disgust as the overall emotion of the face since the disgust emotion has a type-1 feature in the upper face which

Table 3.6 Priority table inside the LogicMax layer.

Lower face CNN output class	**Upper face CNN output class**	**LogicMax output**	**Features**
Neutral/ Disgust / Happy	Fear	Fear	Fear has strong features on upper face
Anger	Fear /Sad/ Surprise/ Happy	Anger	Anger has type-2 feature lip tightener on mouth
Fear/ Sad/ Disgust/ Anger	Neutral	Neutral	Lack of emotion is seen on upper face for Neutral and Happy emotions
Sad	Fear /Sad/ Surprise/ Happy	Sad	Sad has many type-3 features on the upper face
Happy	Sad/ Neutral/ Surprise Fear/ Anger	Happy	Happy has type-1 feature Lip corner puller on lower face
Surprise	Sad/ Neutral/ Anger Fear/ Happy	Surprise	Surprise has type-2 feature Jaw drop on lower face
Sad/ Neutral/ Surprise Fear/ Happy/ Anger	Disgust	Disgust	Disgust has type-1 feature Nose wrinkle on upper face

is very unique when compared to other features but sadness doesn't have type-1 feature in the lower face.

The basics of types - 1,2,3 help in the construction of the priority table. The priority table contains a set of conditional statements for picking the most appropriate emotion of the entire face by analyzing the emotions found on the lower and upper face regions. The priority table has significantly boosted the accuracy of the model during the 10-fold cross-validation process. If the lower CNN and upper CNN output mismatch and do not possess the combinations shown in Table 3.6., then the upper face CNN model output is considered as the final output class for the entire face.

Table 3.7 Comparison of FER accuracy and other parameters on different models for CK+ dataset.

‘	Authors	Method	Testing Procedure	Data Selection	Number of Classes	Performance (%)
CK+	Zhao et al. [69]	The Peak-Piloted Deep Network (PPDN) , 2016	10-fold cross validation	last three frames of each sequence (near to peak expression)	6	97.3%
	Siyue Xie and Haifeng Hu [70]	Facial expression recognition with FRR-CNN, 2017	10-fold cross validation	last three frames of each sequence (near to peak expression)	6	92.06%
	Jung et al. [71]	Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition, 2015	10-fold cross validation	Not mentioned	7	97.2%
	Sherly Alphonse and Dejey Dharma [72]	Novel directional patterns and a Generalized Supervised Dimension Reduction System (GSDRS), 2019	10-fold cross validation	last three to four frame of each sequence (near to peak expression)	7	97.71%
	Yang et al. [73]	Facial expression recognition by de-expression residue learning, 2018	10-fold cross validation	last three frames of each sequence (near to peak expression)	7	97.3%
	The Proposed work	A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition	10-fold cross validation	last three to four frames of each sequence (near to peak expression)	7	98.62%

Table 3.8 Comaparision of FER accuracy and other parameters on different models for JAFFE dataset.

‘Dataset	Authors	Method	Testing Procedure	Data Selection	Number of Classes	Performance (%)
JAFFE	M.K.Mohd Fitri Alif et al. [74]	Fused convolutional neural network for facial expression recognition, 2018	10-fold cross validation	All the images in the dataset	7	83.72
	Caifeng Shan et al. [75]	Facial expression recognition based on Local Binary Patterns: A comprehensive study	10-fold cross validation	All the images in the dataset	7	81%
	Zhao et al. [76]	Facial Expression Recognition via Deep Learning, 2015	10-fold cross validation	All the images in the dataset	7	90.95%
	The Proposed work	A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition	10-fold cross validation	All the images in the dataset	7	94.86%

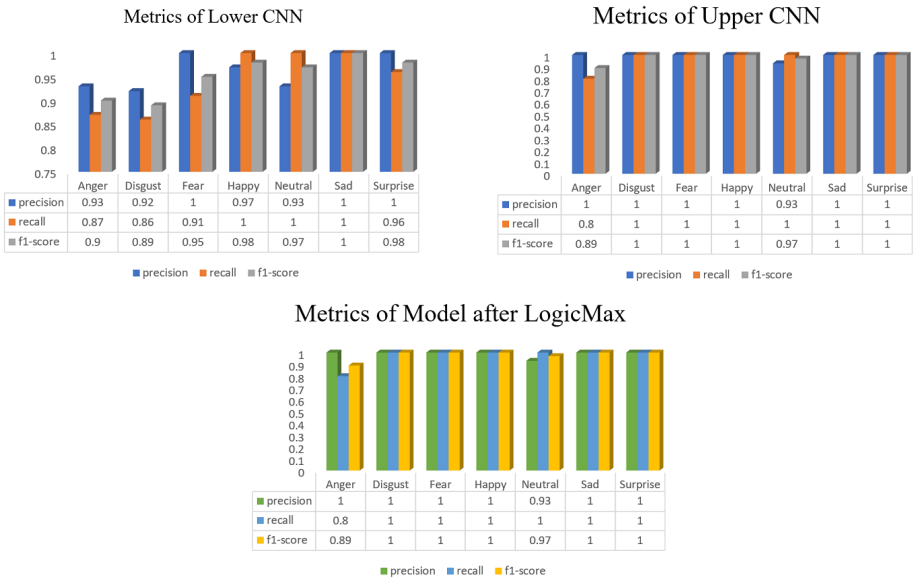

Figure 3.14 Classification metrics of the model on CK+ test data

3.4 Experimental Results and Discussion

In this section, we discuss the facial expression datasets, method of testing the model, results, performance comparison of our model with other significant FER models, and different metrics for evaluation. We then provide filter visualization of the CNN models in Fig. 3.6.

3.4.1 Databases

We have used the two most popular facial expression databases extended CohnKanade database (CK+) and the Japanese Female Facial Expression (JAFFE) database, for testing the model performance.

1. **CK+ database** [77]: The extended Cohn-Kanade, widely known as CK+, is a facial expression dataset for the classification of action units and facial emotion recognition. The dataset has posed as well as non-posed expressions. The Extended CohnKanade (CK+) dataset consists of 593 sequences across 123 different subjects. Considering the most appropriate method, most of the FER models have taken the last three or five frames of the sequence and used them for image-based facial expression recognition. Each sequence in the database contains frames varying from

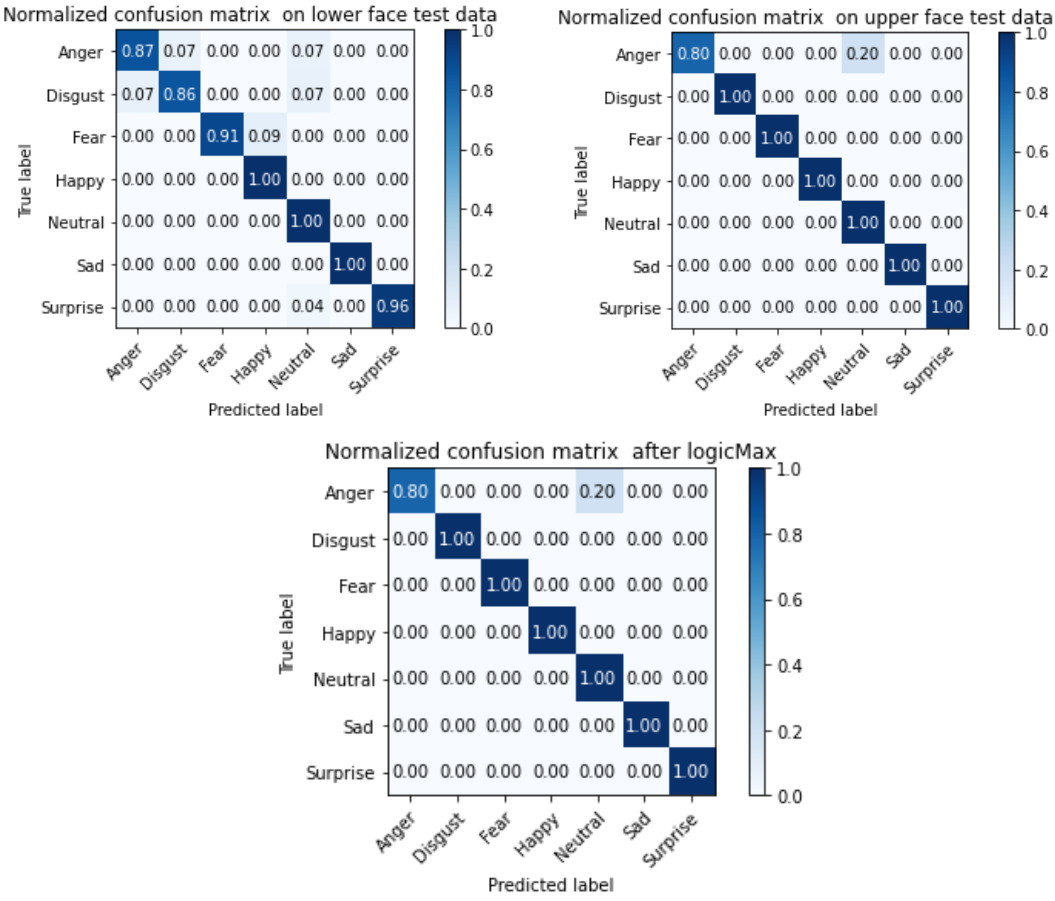

Figure 3.15 Normalized confusion matrices of CK+ test data during 10-fold cross-validation

10 to 60, and in every sequence, the frames are captured such a way that there is a shift in expression from a neutral to the peak intensity of specific emotion. Among the given sequences, only 327 sequences with 118 subjects have the expression labels of anger, contempt, disgust, fear, happiness, sadness, and surprise based on the Facial Action Coding System (FACS). In this work, we have considered the last three to four frames from each labeled sequence for classification. The seven labels taken in this experiment for expression classification are anger, disgust, fear, happiness, neutral, sadness, and surprise. A total of 1479 images are derived from the labeled sequences. The process of extracting two parts of a face is achieved using the dlib library. The images are split into two halves to get the upper and lower face using the dlib library. For the lower face and upper CNN models, 1331 images are used for training the model and 148 images for testing the model using the 10-fold cross-validation process. The emotions predicted by the two CNN models are given as input to the logicMax layer. If there exists a mismatch in the expression

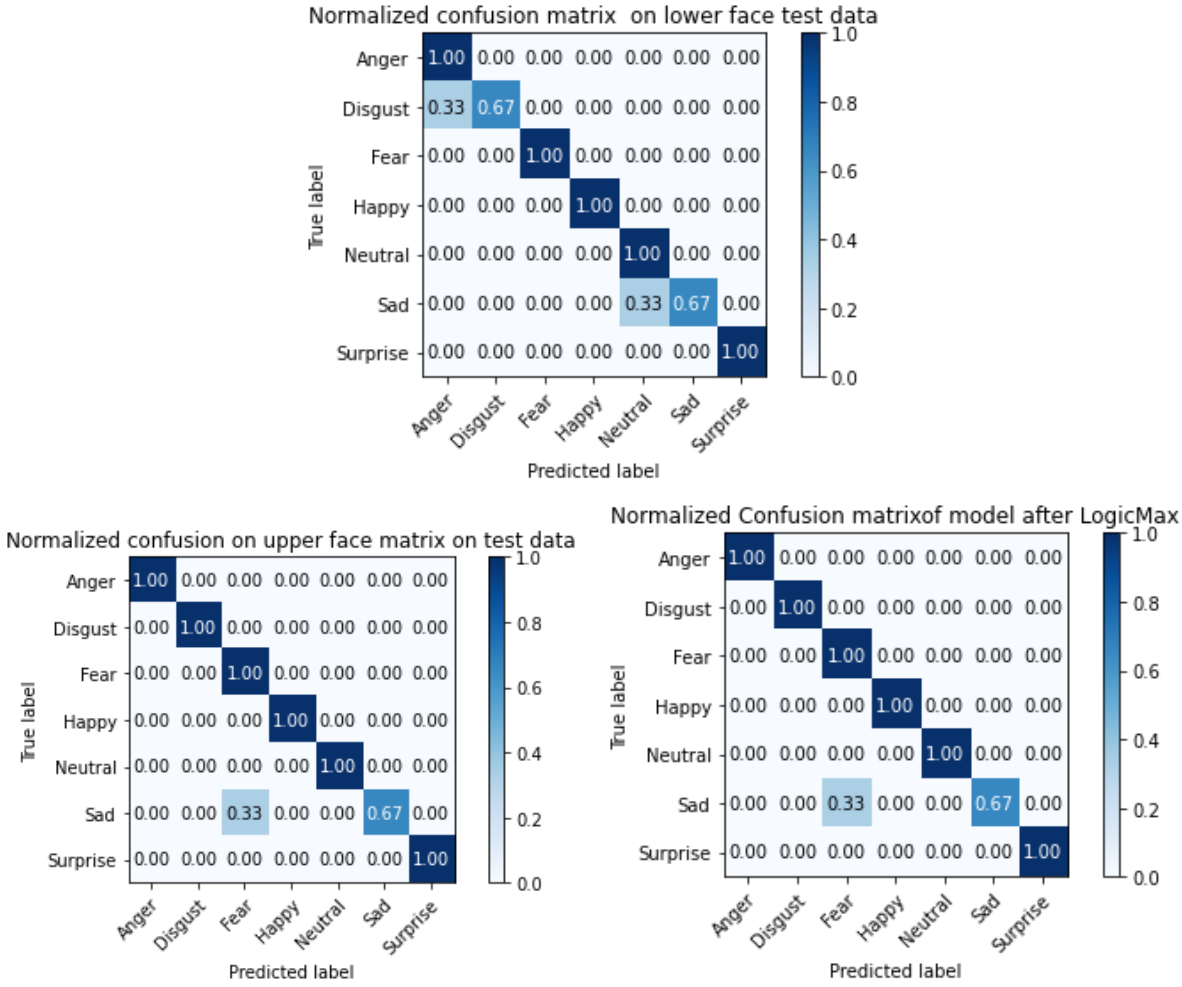

Figure 3.16 Normalized confusion matrices of JAFFE test data during 10-fold cross-validation

classification between the two CNN models then, the LogicMax predicts the final emotion by applying the rules set in the priority table as discussed in the LogicMax and priority table section 3.3.5.

2. **JAFFE database** [78]: Japanese Female Facial Expression has a total of 213 samples, which are posed expressions taken from ten Japanese female subjects. Each subject in the dataset has nearly three to four images of six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and 1 image of neutral expression. This dataset, unlike CK+, has few images for each expression. Data Augmentation plays an important role in this dataset as it could help to extend the number of training samples. We have taken the entire 213 images in the dataset for training and testing the CNN models. The upper face CNN and the lower face CNN models are trained with 192 images and tested with 21 images in each fold during the

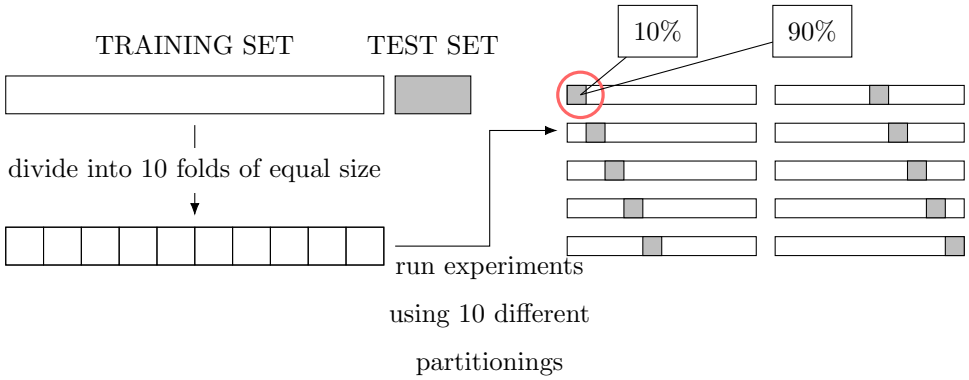

Figure 3.17 k fold validation process, (k=10)

10-cross validation process.

3.4.2 Classification Metrics

The important evaluation metrics of the FER model discussed in this work are Accuracy, Precision, Recall, and F1-score. Let TP represents True Positives, FP represents False Positives, FN represents False Negatives, and FP represents False Positives.

1. **Accuracy**: Accuracy (Acc) is useful in evaluating model performance. However, when there exists a class imbalance problem, it is necessary to consider other important metrics like precision and recall.

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.3}$$

2. **Precision**: The precision (P) highlights the ability of the model to pick the desired class. P depends on TP and FP. False Positives are the number of predictions the model misclassifies as positive when the true label is negative.

$$P = \frac{TP}{TP + FP} \tag{3.4}$$

3. **Recall**: Recall (R) is the other classification metric that conveys the ability of the model to predict all the classes of interest in a dataset. R depends on TP and FN. FN is the number of predictions the model misclassifies as negative when the true label is positive.

$$R = \frac{TP}{TP + FN} \tag{3.5}$$

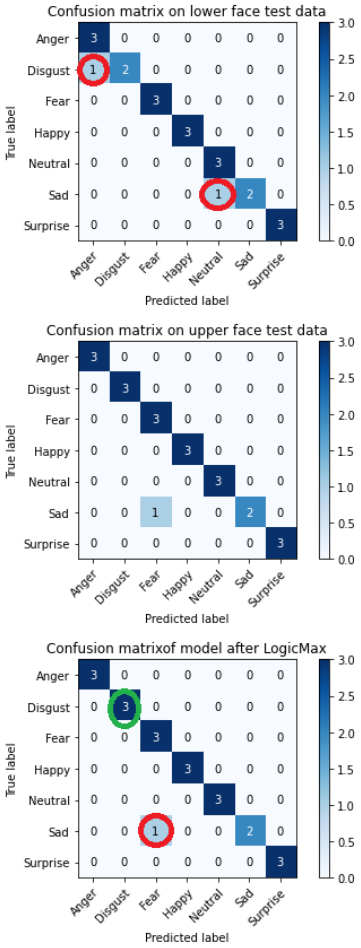

Figure 3.18 Confusion matrices of JAFFE test data during 10-fold cross-validation

4. **F1 Score**: It is necessary to maintain good precision and recall for any model. The goal of a good classifier is to pick the correct class without any mistake (precision) and, at the same time, pick as many as correct classes (recall). A good trade-off is to be maintained between precision and recall. F1 score provides a decent blend of two metrics recall, and precision. F1 score is the harmonic mean of recall and precision.

$$F1\ Score = 2 * \frac{P * R}{P + R} \tag{3.6}$$

3.4.3 K-Fold Validation Process: Testing the Model Performance

Testing in machine learning and deep learning is a fundamental process in evaluating the performance of the model. The popular validation techniques mentioned in the literature are the hold out method, k-fold cross-validation and leave-one-out cross-validation.

In this work, we have used the k-fold cross-validation procedure since it is a widely used evaluation technique in various state-of-the-art methods.

In the k-fold cross-validation process, the total data is randomly partitioned into k equal-sized parts as shown in Fig. 3.17. In these k parts, one part is retained as the validation data for testing, and the other (k - 1) parts are used for training the model. The cross-validation process is then repeated k times, with each of the k folds used exactly once as the validation data. This process results in k individual results, and the scores are then averaged to produce a single estimation. Each sample is used for validation exactly once, which reduces the bias on the data. The value of k is arbitrary. The 10-fold cross-validation is commonly used in many FER models for the evaluation process. The upper and lower face parts are partitioned from the selected data samples and given as input to the respective CNN models. In this research work, we have used the scikit learn's [79] K-Folds cross-validator to split the dataset into k (k=10) consecutive folds with a shuffle. The 10 folds are created such that each fold has nearly 10% of the total data samples for testing. The Table 3.3. provides information about the number of training and testing data samples used in upper and lower face CNN models. The training samples are shown in the Table 3.3. are used to train the CNN models, and the testing is done using the two CNN models and logicMax layer. There is no need for logicMax during the training process since the use of logicMax is required only during the testing phase.

3.4.4 Evaluation of Model Performance and Comparison

In this section, we have shown the results of different metrics to evaluate the performance of the two CNN models. We have performed a 10-fold cross-validation process for evaluating the accuracy scores on both datasets. The confusion matrices of CNN models and the effect of logicMax layer are discussed for both datasets. The important performance metrics accuracy, precision, recall, and F1-Score are shown for CNN models during cross-validation, refer to Fig. 3.14 and 3.8. The importance of the LogicMax can be understood from the confusion matrices shown in the Fig. 3.9 and 3.18. The differences in output during emotion classification between the lower face CNN and the upper face CNN model are corrected by the LogicMax layer, as seen in confusion matrices shown in the Fig. 3.9 and 3.18. A typical example of the advantage of Logicmax is seen in the

confusion matrices, refer to Fig. 3.8. Few samples under disgust are misclassified in the lower CNN model, but due to the unique action unit of disgust on the upper face, the sample gets correctly classified in the upper CNN model, and using the logicMax layer the correct class is selected by the model. The normalized confusion matrices are shown in the Fig. 3.15, 3.16. for understanding the model. The logicMax on analyzing the outputs from the CNN models predicts the correct class. Emotions like disgust and Sadness which are usually hard to recognize have achieved good accuracy using the proposed algorithm. Emotions of happiness, surprise, and disgust have scored good accuracy in both CK+ and JAFFE datasets. It is important to perform a comparison with other models under similar test conditions. The important test conditions that have to be maintained are the number of samples taken from the dataset for training and testing, the number of classes (emotions) that the model can classify, the method of validating the test data, and the iterations used for validating the data. The overall accuracy of the 10-fold cross-validation process on the CK+ and JAFFE datasets is 98.62% and 94.86% respectively. The proposed model is compared with another state-of-the-art model which has used similar test conditions, refer to the Table 3.7 and 3.8.

The problem faced during the training of our model on the CK+ database is the class imbalance issue. In the CK+ dataset, emotions like happiness and surprise have more labels when compared to the labels of disgust and sadness (refer to Fig. 3.1). The class imbalance issue can be partially solved by creating more samples through data augmentation. However, the additional images created during the data augmentation are useful only during the training process but not used in testing. In recent years, the class imbalance issue is solved by generating synthetic data through GANs.

3.4.5 Filter Visualization

It is important to visualize the filters in our CNN model. Each CNN tries to capture low-level features like edges and corner points in initial layers. In the VGG network Fig.3.5, it is clear that as we move deeper into the network, there is a decrease in the kernel size from the 2nd convolution layer (224x224) to the last convolution layer (14x14). The increase in the feature maps is also seen in the architecture from 64 feature maps in the initial convolution layers to 512 feature maps in the last convolution layer. Since it is

difficult to visualize all the filters from each layer, we have shown the first 64 convolutional filters in the first layer of our model in Fig. 3.6.

3.5 Concluding Remarks

The classification of facial expressions using FACS and LogicMax has improved the accuracy rates on CK+ and JAFFE datasets. The performance of the model and other parameters are compared with other state-of-the-art techniques, and the proposed model achieved a good accuracy score. This work improves the precision of classifying emotions like Happiness, Disgust, and surprise by implementing a dual CNN architecture. The LogicMax analyzes the predicted emotions found on the upper and lower face and decides the final class by selecting the most appropriate emotion. The proposed work can be extended by using other correlation methods on action units. In the future, the proposed model is planned to be implemented on embedded hardware platforms.

Chapter 4

A Novel Multi-Feature Fusion Deep Neural Network using HOG and VGG-Face for Facial Expression Classification

4.1 Introduction

The proposed design in this section uses a combination of self-learnt CNN and hand-engineered features. Since the model considers two different features derived using two unique algorithms, the efficiency of the FER is improved. A dual-input deep neural network is finally evaluated on three popular facial expression datasets using the five-fold and ten-fold cross-validation technique. In the following sections, the architecture and advantages of multi-feature fusion are discussed in a detailed manner. The existing multi-feature fusion FER are also compared in this research work. The proposed model uses a Histogram of oriented gradients and VGG-Face-based features. The combination of two unique features extracted through two different techniques has outperformed other FER models. In overcoming the overfitting issue current models are generally trained on huge datasets but when the training samples are sparse the current DNN based models fail to overcome overfitting issues. Instead of increasing the samples, our proposed model increases the number of features from the same limited samples. Due to the increase in the features the DNN models tend to learn more when compared to training on only deep learnt features. In the following section, we discuss how multi-feature fusion models

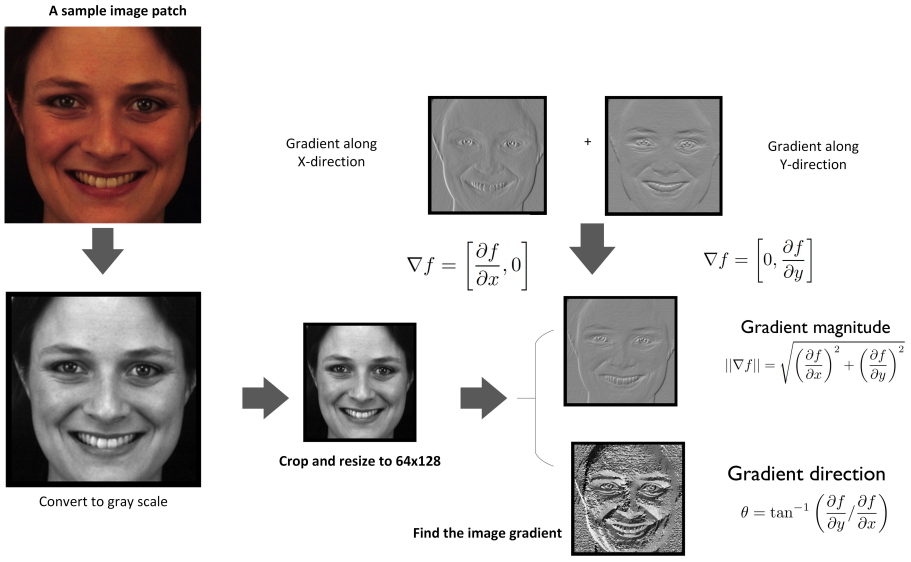

Figure 4.1 The Procedure to calculate the magnitude and direction of the gradient

possess robustness when compared to single-featured models.

4.2 Extracting HOG and CNN features

4.2.1 Extraction of HOG [1] feature descriptors

The proposed FER model is a multi-input deep neural network that considers HOG feature descriptors as one of the input sources.

Can HOG descriptors capture micro facial expressions? HOG features count the localized gradient changes around each pixel, and it is also intuitively beneficial for modelling the structure of the facial muscles through edge analysis. Carcagnì et al. [80] have performed various experiments on selecting an appropriate hand-engineered feature for designing a FER model. The experiment considers four important features that are Local Binary Pattern (LBP) [81], Spatial Weber's Local Descriptor (SWLD) [82] [83], Compound Local Binary Pattern (CLBP) [84] and HOG features. The experiment has revealed the performance of the four features in classifying six facial expressions. Cohn -Kanade (CK+) dataset was used to evaluate the performance, and the results of 10-fold cross-validation accuracies are shown as follows (LBP 91.7%, CLBP 92.3%, SWLD 86.5%,

and HOG 95.8%). These findings revealed that HOG descriptors better model facial muscles than texture-oriented descriptors and are well suited to describe facial expressions. The experiment also shows that the texture-based features are unable to explain facial deformation between different expressions. The facial expressions contain various micro-expressions, which are usually facial muscle deformations. The HOG features are well sensitive to object deformations, and hence it is well suited for facial expression recognition. Chen et al. [85] has performed facial expression detection based on facial components and HOG. In the experiment, the authors used HOG to model various facial components like mouth and eyebrows. The results show that HOG has performed better than Gabor and local binary pattern features. Thus, in this work, we have included HOG features since has shown better performance than other hand-engineered features.

The method for extracting HOG features on faces is discussed in this section. The following measures illustrate the whole technique for extracting the characteristics of HOG:

1. Patch creation: The image provided is sliced under certain conditions into separate patches. The size of the patch is subjective, and patches typically have a fixed aspect ratio (1:2). Examples of sizes for patches are (64x128) and (100x200). The HOG descriptor for each patch obtained is computed. These image patches contain several important facial landmark points that are essential in detecting expressions

2. Calculation of the image gradient on 8x8 cells: The collected patch is split into 8x8 cells. For each evaluation of the 8x8 cell, gradient magnitude and the direction of the gradient are computed, refer to Fig. 4.1. The gradient orientation is scaled between 0-180 degrees. The 0-180 scale is divided into bins, the number of bins is arbitrary (mostly 8 to 12). The corresponding magnitudes are measured at different pixels and are positioned in each bin. The bin-histogram is transformed to a 9x1 vector (if orientations are 9), and each cell in the patch has a 9x1 vector using the same process, refer Fig. 4.2. The gradient variations, both in magnitude and direction, that occur near the important facial landmark points are calculated in this step.

3. Block Normalization: To create a 16x16 block, the 8x8 cells are merged. A mixture of four 8x8 cells creates a 16x16 block. The 16x16 block comprises four histograms,

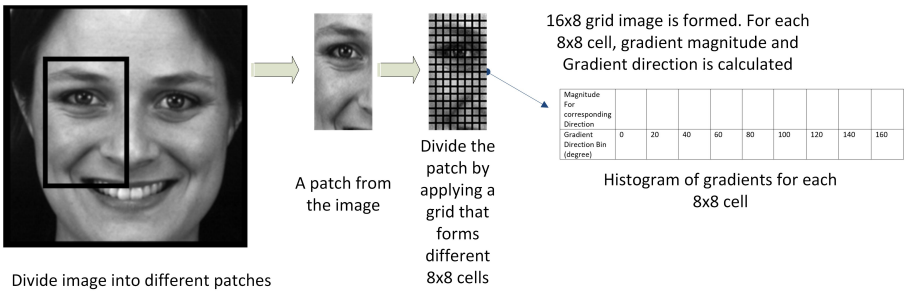

Figure 4.2 Creating a grid of size 16x8 and obtaining the histogram of gradients for each 8x8 cell

and normalizing these histograms is the next step. The gradients are usually dependent on the intensity of the pixels. If the picture has some lighting differences, it will change the amplitude of the gradients. Normalizing can render the gradients independent of variations in lighting. The four histograms (9x1) are concatenated to form a 36x1 normalized vector. This step provides better in-variance to noise that is usually observed in facial regions. The crucial drawbacks of FER models, such as illumination and shadowing effects on facial landmark regions, are avoided using the block normalization technique.

4. HOG feature descriptor: Finding a feature descriptor for all image patches is the final phase. There are seven horizontal and 15 vertical blocks on each patch, making 7 x 15 = 105 blocks inside a patch. Each block has a 36x1 dimensional vector. The concatenation of all 105 block vectors in a patch result in 36x105x1 = 3780x1 dimensional vector. In this way, the feature descriptors on the given image of all patches are obtained as shown in the visualization of HOG features in Fig. 4.3. The feature descriptors contain overlapping cells of facial regions that contain crucial patterns of micro expressions.

4.2.2 Convolutional Neural Network

A standard convolutional neural network mainly involves applying convolution operations, imposing activation functions, and pooling.

The image can be directly given as the input to the CNN. The model applies filters

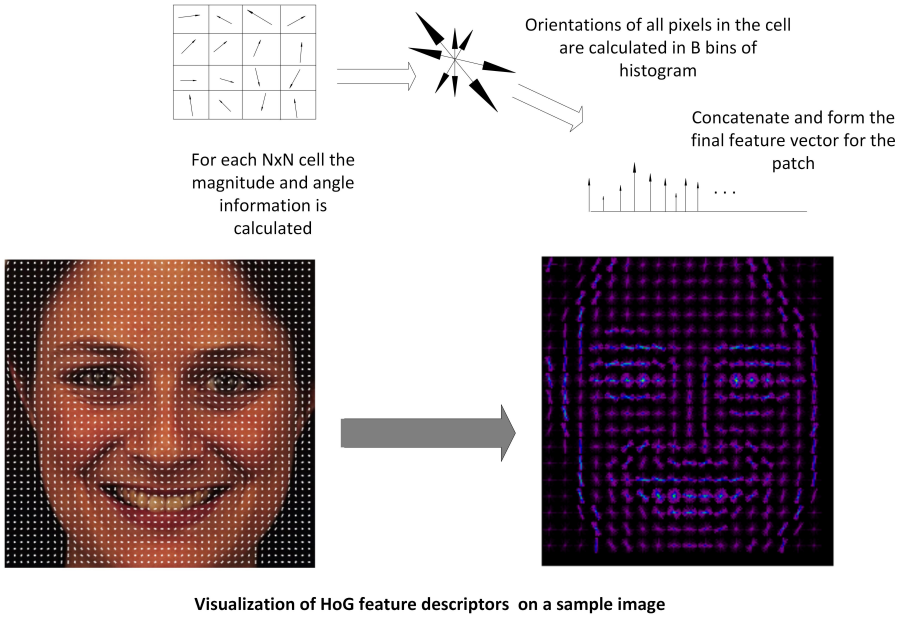

Figure 4.3 Visualization of HOG features

to the image, resulting in generating feature maps. The kernel (filter) size, number of filters, and type of padding are the essential hyper-parameters to be chosen. The hyper-parameters determine the depth of the output feature map. Let us consider **I**[j,k] as the input image, **H** is the kernel, and **O** represents the output matrix after convolution. The equation 4.1 provides the method to determine the size of the images after convolution.

$$\mathbf{O}[m,n] = (\mathbf{I} * \mathbf{H})[m,n] = \sum_{j}\sum_{k} \mathbf{H}[j,k]\mathrm{x}\mathbf{I}[m-j,n-k] \tag{4.1}$$

CNN has multiple padding systems in existence. CNN's various padding schemes are valid, full, and the same. The size of the output matrix after convolution is reduced by valid-padding. At the same time, the size of the output matrix is the same as the input matrix with the same-padding. The other significant characteristic of convolution is the method of stride. Striding is the magnitude of shift taken by the kernel during the convolution processing stage. During the convolution process, the kernel slides over the image. The stride number defines the step size of the shift. If we want to reduce the dimensions of the feature map or decrease the repetitive activity, we can increase the stride value. Let n represents image size, f represents filter size, c represents the colour map depth (number of channels), p is the used padding, s represent stride number, and z is the number of filters. The below equation 4.2 shown gives the output feature map

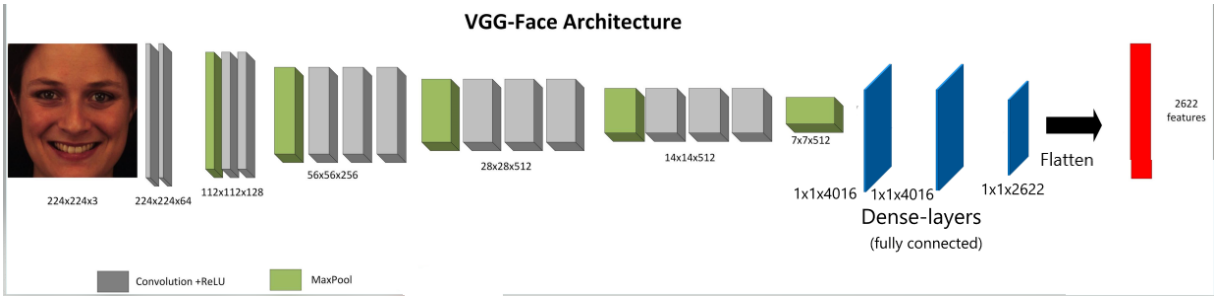

Figure 4.4 Architecture of VGG-Face

dimensions after the convolution of the image with the filters.

$$[n, n, c] * [f, f, c] = [[\frac{n + 2p - f}{s} + 1], [\frac{n + 2p - f}{s} + 1], z] \tag{4.2}$$

The feature maps obtained after the convolution are linked to an activation function. Generally, Rectified Linear Unit (ReLU), F(x)=Max(0,x), is the preferred activation function in computer vision problems. ReLU is also placed after the convolutional layer to provide non-linearity in the model. A special layer called the pooling layer, which down-samples the features is introduced after the convolution. The pooling layer reduces the feature map's size by shrinking the image and retaining critical features simultaneously.

Max-pooling and average-pooling are typical CNN pooling strategies. The convolution-ReLU-max-pooling blocks are repeated to produce the few most efficient features from an input image. Finally, through the use of a flatten layer, the 3-dimensional feature maps are translated to 1-dimension. A softmax activation function that produces probabilities (0 to 1) for the respective output class is attached to the final fully connected layer. VGG-Face, shown in Fig. **??**, is a typical CNN model used in this work.

4.2.3 Advantages of Combining HOG and VGG-Face Features

1. The HOG features capture micro-expressions at important landmark points such as eyes, eye-brows, and mouth. These low-level features such as edges and corner points are essential to distinguish emotions on the face.

2. The generalization of CNN is improved by using a fine-tuned CNN model instead of training the CNN model on smaller datasets. Fine-tuning is a transfer learning

Table 4.1 Comparison of popular pretrained CNN models

Model	Authors	Year	Advantages	Applications and Datasets tested
AlexNet [86]	Krizhevsky et al.	2012	Application of Rectified Linear Units (ReLUs) as activation functions	Object Detection ImageNet [33]
VGG-16,19 [87]	Simonyan et al.	2014	Implementation of deeper convolutions through stacking uniform convolution layers	Object Detection ImageNet
VGG-Face [88]	Parkhi et al.	2015	Compution of VGG-Face CNN descriptors that model human faces	Face Recognition Evaluated on exclusive face related datasets like Labeled Faces in the Wild [89] and YouTube faces [90]
Inception V1 [91]	Szegedy et al.	2015	Implementation of convolutions with different kernel size filters through addition of inception modules	Object Detection ImageNet
Res-Net 50 [92]	Kaiming et al.	2015	Application of skip connections and Batch normalization	Object Detection ImageNet
MobileNet	Howard et al.	2017	Introducing streamlined architecture that uses depth-wise separable convolutions	Efficient model for mobile and embedded vision applications Evaluated on ImageNet, Fine Grained Recognition, Large Scale Geolocalizaton and Face attributes

approach that focuses on preserving and transferring information obtained while addressing one problem to a different but related problem. Since CNN's are made up of several layers and a large number of parameters, the usage of pre-trained weights, which are calculated on larger databases through fine-tuning, should help prevent the problem of overfitting. When trained on smaller datasets, the CNN models tend to capture patterns that are specifically related to the dataset. This behaviour can make the model over-fit and difficult to generalize on other external data. When the CNN models use fine-tuning, we can reuse the filters that already possess rich weights that show strong discriminative ability. The other advantage of fine-tuning is the reduced count of epochs required to converge the error. Zavarez et al. [93] have compared the differences with and without fine-tuning the VGG-Face model, and the results demonstrated that fine-tuning the VGG-Face that has already been trained on a similar domain is superior to training from scratch. The authors have also performed cross-database facial expression recognition and the fine-tuned VGG-Face has given state-of-the-art results on well-established datasets

like CK+ [94], MMI [95], The Radboud Faces Database (RaFD) [96], KDEF [97], JAFFE [98], and AR Face [99].

3. The normalization step in extracting HOG descriptors improves the performance of the FER model by introducing in-variance to illumination, shadowing, and edge contrast effects. The normalization of cells across larger regions such as blocks can make features in-variant to lighting conditions on the local level.

4. HOG features are relatively constant within the cell rotations and translations, and they are invariant to local geometric and photometric variations. Since critical facial expressions are concentrated on small portions of the face (facial landmarks), HOG features extract better low-level features on the face.

5. VGG-Face, compared to other famous CNN models, is exclusively trained on human faces taken from the Internet Movie Data Base (IMDB) celebrity list. The VGG-Face is originally designed for face recognition and thus can capture important low and high-level features on the face. The Table 6.1 compares various CNN models and their applications, and the VGG-Face model is selected in this architecture since it has proved its efficiency on well-established datasets such as Labeled Faces in the Wild and YouTube Faces.

6. The combination of HOG and VGG-Face can effectively work on collecting important facial traits. The HOG features concentrate on local facial muscle deformations and gather important low-level facial features. The VGG-Face features deal with high-level features by generalizing the local features extracted on faces. The VGG-Face extracts abstract level features from different face patterns. The multi-feature fusion of HOG and VGG-Face also increases the efficiency of the FER since the model depends on two unique features, which are generated using two different methods.

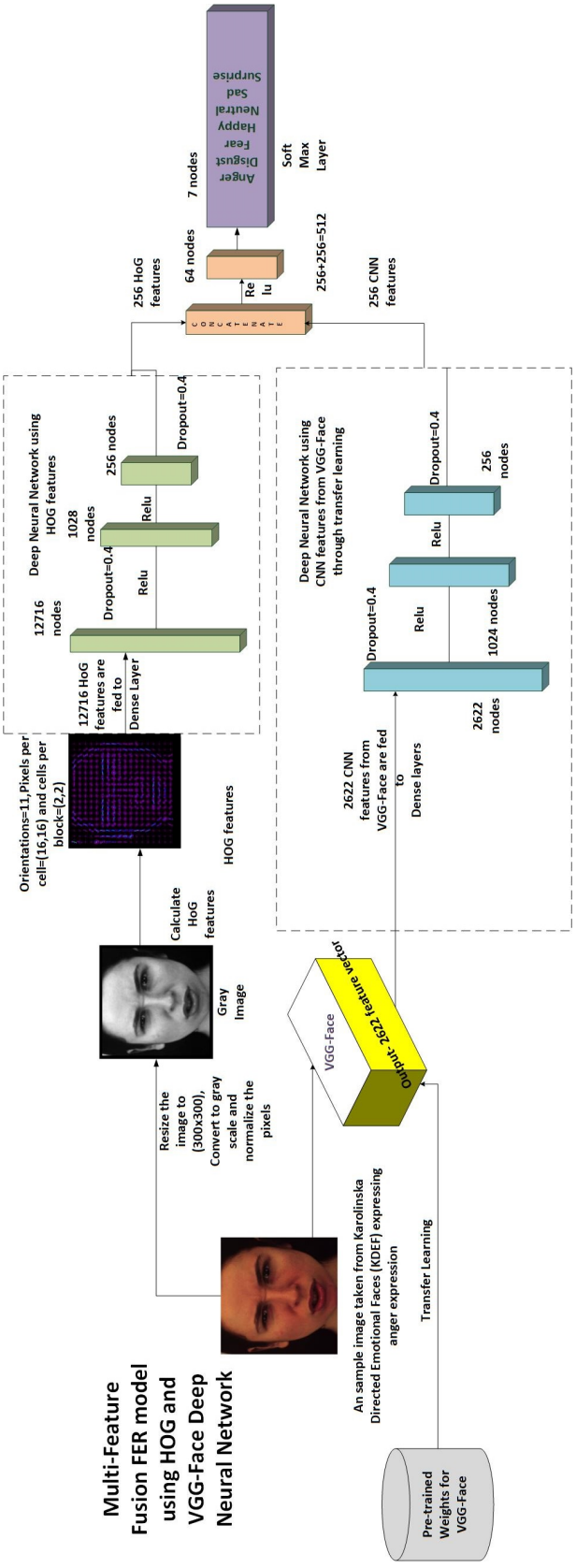

Figure 4.5 The proposed multi input hybrid FER model for facial expression classification

4.3 Design of Multi feature-fusion model

4.3.1 Design Parameters of Multi-Feature Fusion model

The proposed model utilizes both HOG and VGG-Face features to extract features of micro-expressions. It is crucial to choose the best design parameters to increase the efficiency of the FER. The essential parameters for selecting the HOG features are the cell size and the number of orientation bins. Cell size indicates the dimension of the patch used in the single histogram computation. When a large cell size is used, the appearance information of a comprehensive portion of the facial image is compressed into a single cell histogram, and certain features that are important for classification are lost. On the other side, the high-resolution analysis may be performed with small cell size, but this requires the classifier to distinguish between valuable and irrelevant extracted information. It may not be capable of doing well. Thus, selecting an appropriate cell size of (16,16) has given better efficiency for our model. The number of orientation bins is the other parameter chosen, corresponding to the quantization stages of gradient information. A minimal number of orientations may result in some information loss and, as a result, a decrease in FER performance. Many quantization levels, on the other hand, might spread the information over the bins and hence decreases the FER performance. The research on optimizing the HOG features has shown that orientation bins of 8 to 12 have significantly interpreted HOG features in analyzing the facial muscles. The comparison of different HOG attributes with accuracy rates is shown in Table 4.2.

The designed FER model is capable of classifying facial expressions into seven emotions. Angry, disgust, fear, happy, neutral, sad, and surprise are the seven emotions that the designed FER model in this model can identify. The other hyper-parameters of the multi-feature fusion model are explained in Table 4.3. The prepossessing steps of the proposed model include face extraction using the viola-jones method [100], pixel normalization, and image- resizing. Data-Augmentation is not used in the proposed method since the imported weights of VGG-Face are frozen and already trained extensively on human faces.

Table 4.2 HOG feature parameters vs Accuracy rates on KDEF and CK+ dataset

Orientations	Pixels per cell	Cells per block	Accuracy HOG+SVM	Number of features
11	(16, 16)	(2, 2)	86%, 89%	12716
11	(8, 8)	(2, 2)	85%, 87%	57024
11	(4, 4)	(2, 2)	85%, 83%	240944
7	(8, 8)	(2, 2)	84%, 85%	36288
7	(16, 16)	(2, 2)	84%, 87%	8092
7	(4, 4)	(2, 2)	84%, 82%	153328
9	(8, 8)	(2, 2)	85%, 86%	46656
9	(16, 16)	(2, 2)	86% ,88%	10404
9	(4, 4)	(2, 2)	82%, 81%	197136

4.3.2 The architecture of the proposed FER model

Input-I: The designed FER model has two inputs. The first input is obtained from VGG-Face [88] architecture, the inputs to the VGG-face consist batch of RGB images of faces displaying expressions of various emotions. The pixels of the input image is normalized and resized to the size of (224x224x3). As shown in Fig. 4.5, the images are reshaped and given as input to the VGG-Face model. In Fig. 4.4, a sequence of (conv-relu-pool) layers is applied to the input images, and the resulting feature maps are flattened to generate 2622 parameters. The 2622 features are self-learnt features from the VGG-Face. The weights of the VGG-Face are imported through transfer learning. Transfer learning is the technique of reusing a trained model to a new problem. The VGG-Face architecture is trained and tested extensively with massive image data extracted from datasets like Internet Movie Data Base (IMDB) celebrity list, LFW [89], and YTF datasets [90]. It is wise to import the VGG-Face weights instead of training with a smaller dataset. The weights of the VGG-Face model are imported from the data provided by the Visual Geometry Group, University of Oxford [88]. The provided weights from the source are in MatLab format. MatConvNet is a MatLab toolbox for CNN, and the provided weights are MatLab compatible. In this work, the weights of VGG-Face are converted from MatLab format to Keras compatible using machine learning libraries. The detailed procedure of

Table 4.3 Details of Hyperparameters considered for the proposed model

Type of Feature/Model	**Hyper Parameter**	**Value/Function Considered**
HOG	Cell Size	(16.16)
	Number of Orientation Bins	11
Multi Feature Fusion Model	Input Image Size Branch 1 (HOG)	300x300
	Input Image Size Branch 2 (VGG-Face)	224x224x3
	Number of Hidden Layers Branch 1 (HOG)	Three
	Number of Hidden Layers Branch 2 (VGG-Face)	Three
	Activation Unit	Rectified linear activation function (ReLu)
	Epochs	Adaptive (Early Stop using Validation loss)
	Early Stopping Patience level	Patience = 8 steps Montoring parameter = Validation Loss
	Batch Size	16
	Drop Out	20 - 50 %
	Optimizer	ADAM
	Validation Split	5%
	Error Function	Categorical Cross Entropy
	Number of Output Classes	Six for Yale-Face and Seven for CK+,KDEF. Emotion Classes Angry, Disgust, Sleepy, Wink, Fear, Happy, Neutral, Sad, and Surprise

conversion is explained by Sefik [101]. Transfer learning saves computational time and helps the researchers concentrate on the actual problem statement. The imported model provides 2622 deep features, and these features are combined with the HOG features in the proposed model. In Figs. 4.18, the visualization of different feature maps is shown. The feature map visualization at different layers explains the details of high and low-level facial patterns.

Input-II: The corresponding input-I images are resized to (300x300x3). The second input to the FER model is the HOG feature- vectors extracted from the input images given to the VGG-Face. The RGB images are transformed to grayscale to achieve HOG properties. The grayscale images are divided into patches, and for each patch, the information about the histogram of oriented gradients is stored in the form of feature vectors. For every input-I image, a total of 12716 HOG parameters or features are identified; refer to Fig. 4.5. The histogramming effect in HOG makes the features transnational invariant and adapts even in lighting changes. The in-variance of lighting and noise can also be tackled by normalizing the histogram vectors.

Concatenation of HOG and VGG-Face features: The obtained VGG-Face features (Input-1) are given to a series of densely connected layers. The dropout layers added in-between the layers ensure the model does not overfit the data. The 256 HOG features and the 256 features of the VGG-Face CNN are concatenated to form a 512 combined feature vector as shown in Fig. 4.5. This combined feature vector has both hand-engineered and self-learnt features. A series of densely connected layers and dropout layers connect the 512 features. A softmax activation function is attached to the FER model's end. The softmax layer generates the probability scores for seven output emotion classes for each input image. The softmax layer converts randomly distributed values to an ordered probability distribution that varies between (0,1). For each input image, the Softmax function generates an output vector containing seven values for seven emotion categories in which each value ranges between (0,1). The emotion class which possesses the highest softmax score is selected as the final output emotion class of the respective sample. In the following segment, the metrics comparison of the hybrid FER model and traditional CNN model is explained.

Figure 4.6 Sample images taken from CK+ dataset

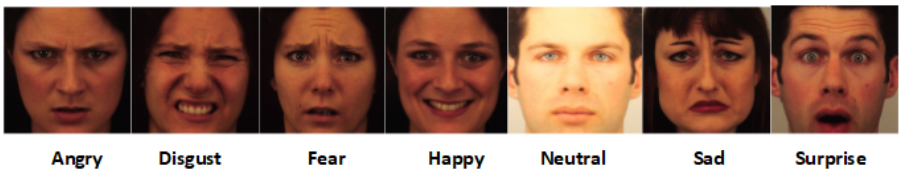

Figure 4.7 Sample images taken from KDEF dataset

4.4 Datasets

We have used three facial expression datasets in our research work.

Extended Cohn-Kanade database [94]: The extended Cohn-Kanade dataset, also known as CK+, includes sequences of images that convey emotion from a neutral state. It contains the descriptions of the action units for peak expression (last frame) images. There are both posed and non-posed facial expressions in the dataset. There are 593 sequences spanning 123 different subjects in the Expanded Cohn-Kanade (CK+) dataset, but it has labels for only 327 sequences. The dataset comprises nearly 10 to 60 frames per sequence. The frames are structured so that the first frame in the sequence displays a neutral expression, and the other frames in the sequence display an emotion shift. The last frame expresses the peak strength of emotion in the sequence. The sample images of the CK+ dataset is shown in Fig. 4.6. In the Extended Cohn-Kanade (CK+) database, only the peak frame of a sequence is fully FACS coded. In the model, the authors Lucey et al. [94] have labeled the CK+ data according to the FACS coded emotion labels. The work explains that the emotion labels assigned to the sequences are validated using the FACS Investigators guide [102] and confirmed by visual inspection by emotion researchers. The authors of CK+ have labelled the emotions according to the FACS coded emotion labels. Since the FACS information was calculated on the expression at the peak phase, only the last frames of the sequences are considered in this research. The

CK+ dataset has sequences (collection of frames) that start with a neutral frame, and the authors considered adding neutral expression class to the model adds robustness to the model since most of the time, faces express neutral expressions. It is important to classify between neutral and emotional faces. We only extracted the last and first frames of all the labelled sequences in the dataset in this experiment. The dataset has a clear class imbalance, refer to Fig. 4.9. The samples of neutral, happy, and surprise are relatively higher than the samples of fear and sad.

The Karolinska Directed Emotional Faces [97]: The Karolinska Directed Emotional Faces (KDEF) database includes 4900 images of human facial expressions with a resolution of 562 x 762 pixels. In this dataset, a total number of 70 subjects are used. All 70 subjects exhibited seven distinct emotional expressions. Each expression is presented at five different angles in this dataset (full left profile, half left profile, straight, half right profile, full right profile). We just focused on straight poses in this proposed work. There are 692 images of seven distinct emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise) in the extracted images. A five-fold and ten-fold cross-validation method is used to evaluate the classification accuracy of the proposed model. The sample images of the KDEF dataset are shown in Fig. 4.7.

The Yale Face Database [3]: The database provides a total of 165 GIF images which include 15 persons (14 males and 1 female). The database consists of 11 images per subject, one for each of the following facial expressions or configurations: center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink. The neutral expressions were obtained by illuminating the face in three positions with a Luxo lamp. The other facial expressions were kept in ambient lighting, and these illumination differences in images make the dataset challenging to classify, refer to Fig. 4.8. In this research work, we have considered six different facial expressions from the Yale face database. The facial expressions of normal, surprised, happy, sleepy, wink, and sad are considered in this work. The total number of images that are extracted from this database is 120. The samples extracted for each class are shown in Fig. 4.11. The dataset is different compared to KDEF and CK+ since faces are subjected to illumination variations. The other interesting new facial expressions, such as wink and sleepy, are also available in this dataset. The dataset is tested using a five and ten-fold cross-validation process on the

Figure 4.8 The Yale database contains 160 frontal face images covering 16 individuals taken under 10 different conditions: A normal image under ambient lighting, one with or without glasses, three images taken with different point light sources, and five different facial expressions [3]

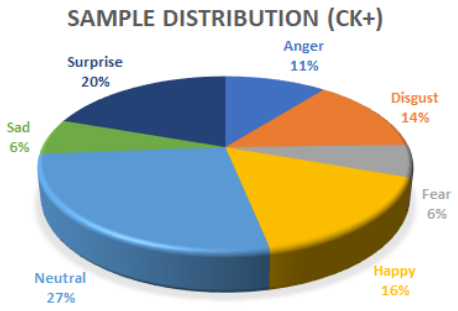

Figure 4.9 Sample distribution of emotion classes in CK+

fine-tuned VGG-Face CNN and the proposed multi-feature fusion model.

4.5 Results and Discussion

4.5.1 Training the proposed FER model with facial expression datasets

The preparation of a FER model using deep neural networks necessitates a GPU. Using Google's Colaboratory (colab), we completed the training of the FER model. Colab provides Nvidia 1xTesla K80 GPU with 2496 CUDA cores. It has a VRAM GDDR5 of 12GB. A single-core hyper-threaded Xeon Processor with a clock speed of 2.3GHz is also

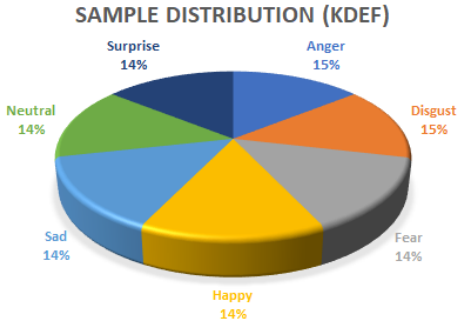

Figure 4.10 Sample distribution of emotion classes in KDEF

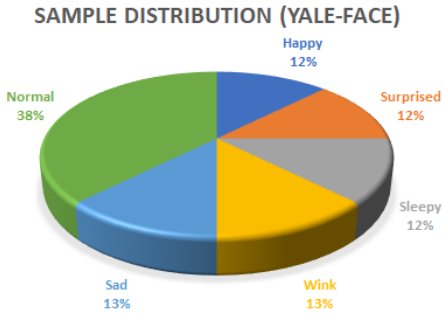

Figure 4.11 Sample distribution of emotion classes in Yale-Face

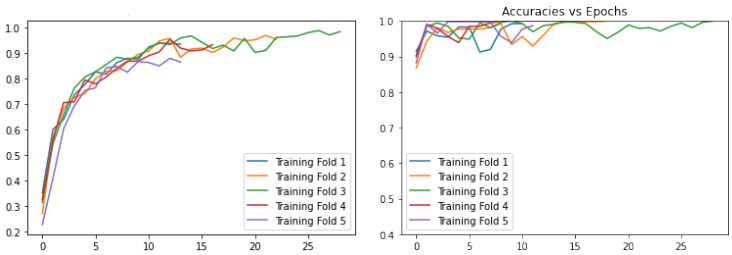

Figure 4.12 Accuracy vs Epochs plot during the training of KDEF data on CNN (left) and the proposed hybrid FER model (right)

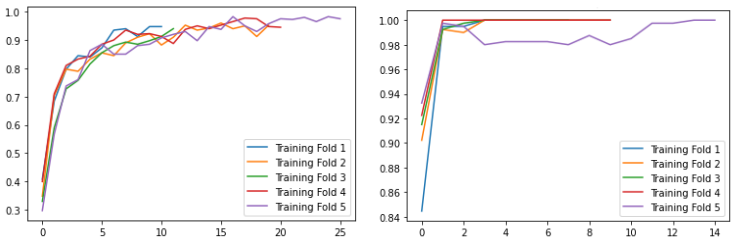

Figure 4.13 Accuracy vs Epochs plot during the training of CK+ data on CNN (left) and the proposed hybrid FER model (right)

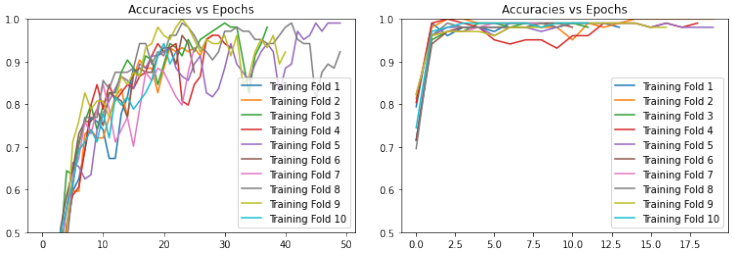

Figure 4.14 Accuracy vs Epochs plots during the training (10-fold-cross-validation) of Yale-Face data on CNN (left) and the proposed hybrid FER model (right)

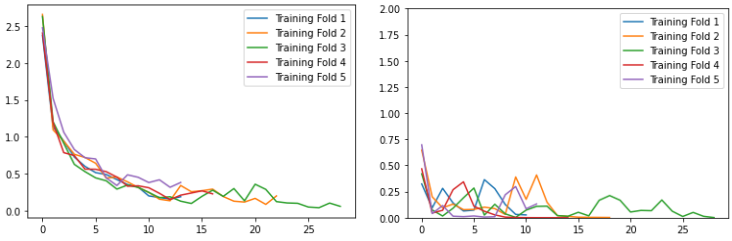

Figure 4.15 The categorical cross-entropy loss vs Epochs plot during the training of KDEF data on CNN (left) and the proposed hybrid FER model (right)

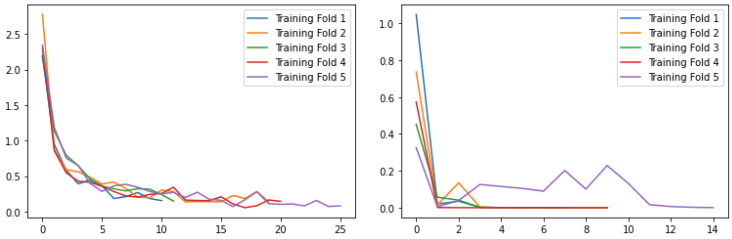

Figure 4.16 The categorical cross-entropy loss vs Epochs plot during the training of CK+ data on CNN (left) and the proposed hybrid FER model (right)

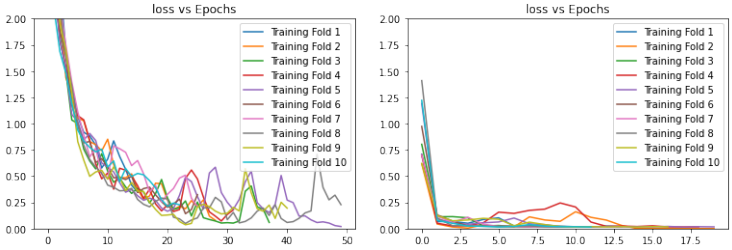

Figure 4.17 The categorical cross entropy loss vs Epochs plots during the training (10-fold cross-validation) of Yale-Face data on CNN (left) and the proposed hybrid FER model (right)

available in Colab. Preprocessing of the dual inputs follows the steps outlined in the previous section. We have used Keras for building the deep neural network architecture. An error function, categorical cross-entropy, is used in this model. Through updating the weights in the neural layers, the error value measured using categorical cross-entropy is minimized. The formula for obtaining the categorical loss is explained in equation 4.4. The weights are updated in each epoch using an optimizer. We used the ADAM [103] optimizer in this experiment because it incorporates the effects of the RMSProp [104] (the ability to deal with non-stationary objectives) and ADAGrad [105] optimizers (the ability to deal with sparse gradients). In ADAM, individual adaptive learning rates for various parameters are calculated using estimates of the first and second moments of the gradients. ADAM possesses various advantages, such as the magnitudes of parameter updates are invariant to gradient rescaling, the stepsizes are constrained by the step-size hyperparameter, and it does not require a stationary objective. The authors in the research [103] also discussed the effectiveness of ADAM in multi-layer neural networks and deep CNNs. The experiments provided in the paper show ADAM to be robust and well-suited to a wide range of non-convex optimization problems in the field of machine learning. The equations and other parameters used in the ADAM optimizer are shown in Table 4.4. The three datasets are divided into five and ten-folds and cross-validation testing procedure is employed in determining the accuracy scores of the model. The training accuracy and loss vs epochs graphs are shown in Figs. 4.12 to 4.17. At the end of the proposed model, the softmax layer produces the predicted emotion values for all input images. The softmax layer is a function that turns random values into a properly ordered probability distribution. The output of the Softmax layer function differs from (0,1). In the proposed hybrid FER model, there are a total of seven/six classes. Let us consider t_i and y_i be the target and the softmax score of the i^{th} class of a sample. The softmax activation function is explained by equation 4.3, for each class i, there exists a softmax score according to equation 4.3. The class with the highest softmax score is predicted as the output class of the respective sample.

$$\textit{Softmax score for each class i= 1 to 7}: f(y)_i = \frac{e^{y_i}}{\sum_j^{N=7} e^{y_j}} \tag{4.3}$$

$$\textit{Categorical Cross entropy error}: -\sum_{i=1}^{N=7} t_i log(y_i) \tag{4.4}$$

Parameter	**Value chosen**	**Role in ADAM**
Epsilon ,ϵ	10^{-8}	preventing division by zero
Learning rate,η	0.001	step size in each iteration
First momentum,β_1	0.9	speed of convergence
Second momentum,β_2	0.99	speed of convergence

Table 4.4 Values of different parameters used in the ADAM optimizer

Table 4.5 Mean Accuracy and standard deviation scores of three models using five and ten fold cross-validation process

Model	**Dataset**	**Accuracy(5-fold cross-validation)**	**Accuracy(10-fold cross-validation)**
HOG+SVM	CK+	88.87% (+/- 1.76%)	89.24 (+/-2.15%)
	KDEF	85.56% (+/- 1.58%)	86.17% (+/-1.23%)
	Yale	56.16% (+/- 0.09)	55.32% (+/- 0.05)
Fine-Tuned VGG-Face CNN	CK+	81.11% (+/- 2.81%)	82.31% (1.12%)
	KDEF	75.84% (+/- 2.53%)	74.45% (+/- 2.31%)
	Yale	48.67% (+/- 2.86%)	50.12% (+/- 1.87%)
HOG + VGG-Face Feature-fusion model	CK+	98.12% (+/- 1.14%)	98.11% (+/- 2.32%)
	KDEF	96.36% (+/- 1.04%)	97.84% (+/- 2.76%)
	Yale-Face	95.26% (+/- 2.5%)	96.67% (+/- 2.15%)

The below equations explain the procedure to update weights using ADAM optimizer. Table 4.4 explains the values of the parameters used in ADAM : $\theta_{t+1} = \theta_t - \frac{\eta \cdot \hat{m_t}}{\sqrt{\hat{v_t}} + \epsilon}$

where

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t},\ \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

and where

$$m_t = (1-\beta_1)g_t + \beta_1 m_{t-1}$$

$$v_t = (1-\beta_2)g_t^2 + \beta_2 v_{t-1}$$

$$g(gradient) = \nabla J(\theta_{t,i})$$

The sum of all outputs from the softmax layer equals one. In Multi-Class classification problems, the targets are one-hot encoded, making only the positive class appear in the categorical loss function.

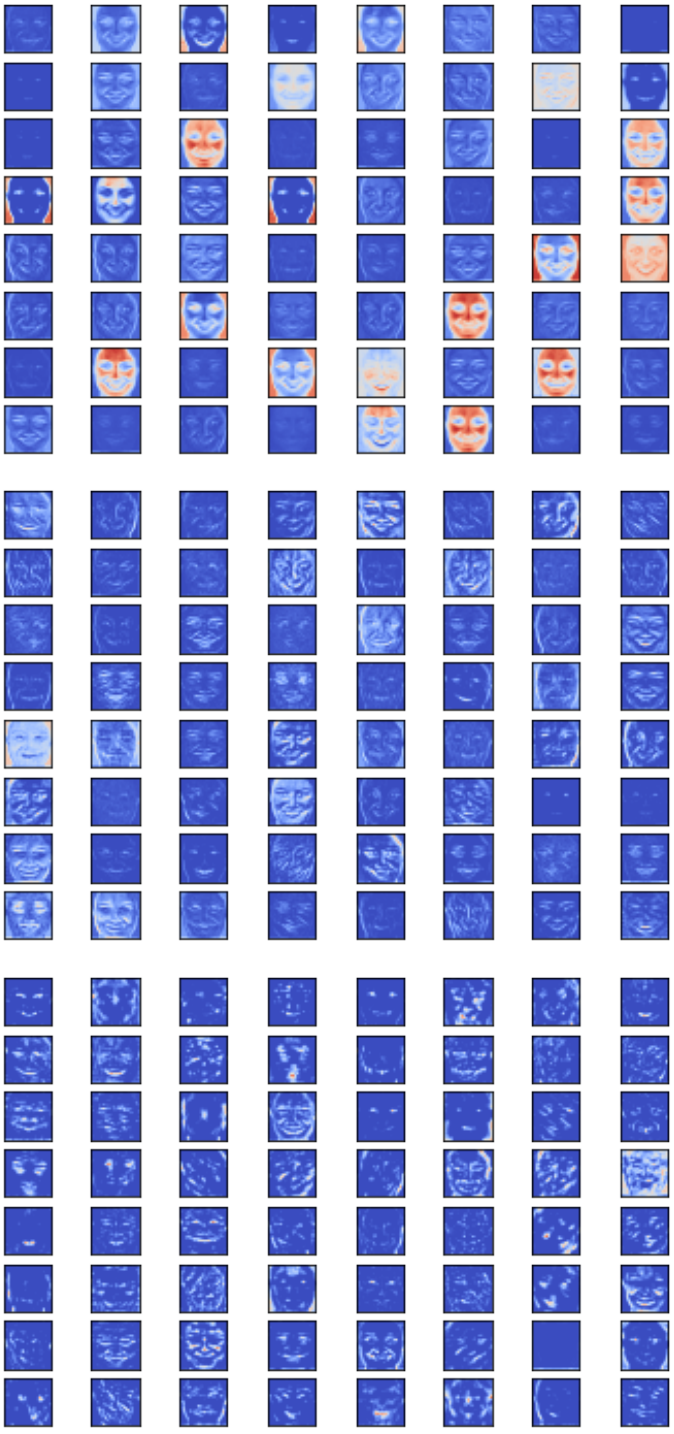

Figure 4.18 Filter visualization of 64 feature maps at different convolutional layers in VGG-Face

4.5.2 Classification Metrics

The model's essential evaluation metrics discussed in this work are accuracy, precision, recall, and F1-score. Let TP represents True Positives, FP represents False Positives, FN represents False Negatives, and FP represents False Positives.

1. **Accuracy**: Accuracy (Acc) is useful for evaluating model efficiency. However, when there is a class imbalance problem, it is essential to consider other critical metrics, such as precision and recall.

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{4.5}$$

2. **Precision**: Precision (P) underlines the ability of the model to select the class of choice. P is based on TP and FP. False Positives are the number of predictions that the model misclassifies as positive when the true label is negative.

$$P = \frac{TP}{TP + FP} \tag{4.6}$$

3. **Recall**: Recall (R) is the other classification metric that conveys the ability of the model to predict all classes of interest in the dataset. R is based on TP and FN. FN is the number of predictions that the model misclassifies as negative when the true label is positive.

$$R = \frac{TP}{TP + FN} \tag{4.7}$$

4. **F1 Score**: Good precision and recall must be preserved for every model. A good classifier aims to choose the correct class without any error (precision) and, at the same time, to choose as many correct classes as possible (recall). A successful trade-off between precision and recall must be preserved. The F1 score offers a good combination of two measures of recall and precision. The F1 score is the harmonic mean of recall and precision.

$$F1\ Score = 2 * \frac{P * R}{P + R} \tag{4.8}$$

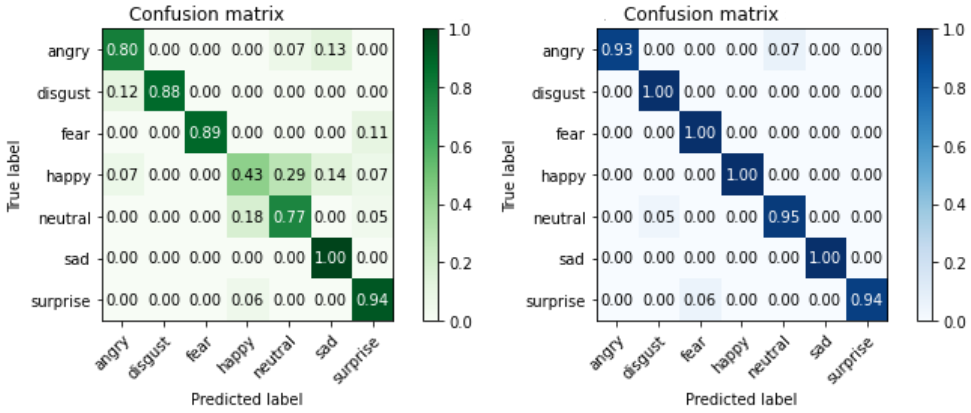

Figure 4.19 The normalized confusion matrix plots of the CNN (left, green) and the proposed hybrid FER model (right, blue) on CK+ test data

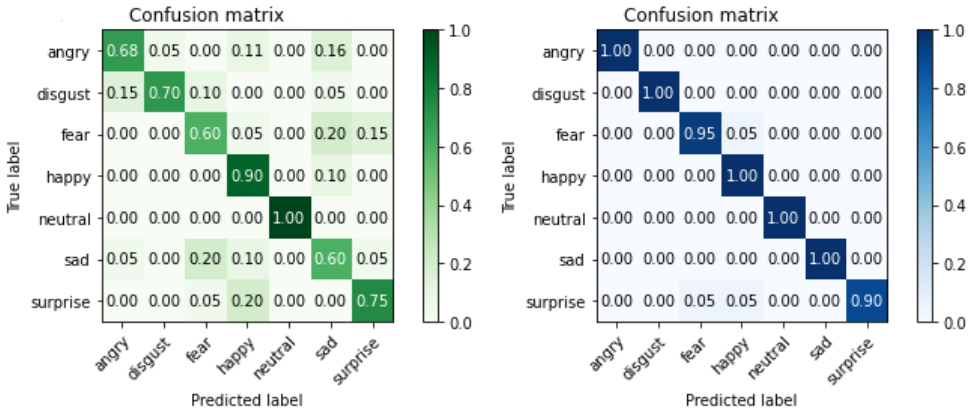

Figure 4.20 The normalized confusion matrix plots of the CNN (left, green) and the proposed hybrid FER model (right, blue) on KDEF test data

4.5.3 Comparison of cross-validation results of the CNN and the proposed model

In this section, the authors have compared the designed multi-feature fusion model with a conventional CNN model. For creating a CNN based FER model, the flattened 2622 features of the VGG-Face, refer to Fig. 4.4, are extended using hidden and dropout layers and finally connected to a softmax layer containing seven classes. The 2622 features are connected with three hidden layers (with dropout) similar to branch-2 of the proposed model, refer to Fig. 4.5, resulting in 256 features. The 256 features are finally connected to a softmax layer of 7 classes. The comparison of fine-tuned VGG-Face CNN model and the proposed multi-feature fusion model explains the improvements achieved in feature fusing CNN and HOG features. The accuracy vs epoch graphs during model training in each fold of the cross-validation phase is shown in Fig.4.12. Compared to the CNN model, the proposed hybrid FER model trains faster during each training fold because it incorporates more details or features, making the deep neural network easy to find more complex patterns of the micro-expressions. The Figs. 4.12 and 4.15 explain the accuracy and loss value plots with epochs. It is observed that the proposed model achieves higher training accuracy scores in a few epochs and attains maximum training capability within a few epochs. The proposed model minimizes the cross-entropy error in less number of epochs when compared to the traditional CNN model, refer Figs.4.15,4.16.

As validation data, we used approximately 5% of the training data. Validation data is beneficial when constructing an impartial model. The 5% validation split aids model generalization and helps prevent the model from overfitting to the training data. The model was also designed with an early stopping technique to avoid over-fitting. The early stopping technique monitors the validation loss in each epoch, and if the validation loss does not improve, the early stopping technique interrupts the model's training. The patience level has been set to 8, which means that if the validation loss does not improve for eight epochs in a row, the model's training must be stopped. The early stopping procedure prevents the model from being over-fitting to the data. It is due to the early stopping procedure that the graphs of the training accuracy plots in Figs. 4.15 and 4.16 have different epochs executed for different training folds. In training the KDEF data, refer to Figs. 4.12 and 4.15, the proposed model shows certain instability in training the data when compared to the CNN model. The effect of a high patience value in designing

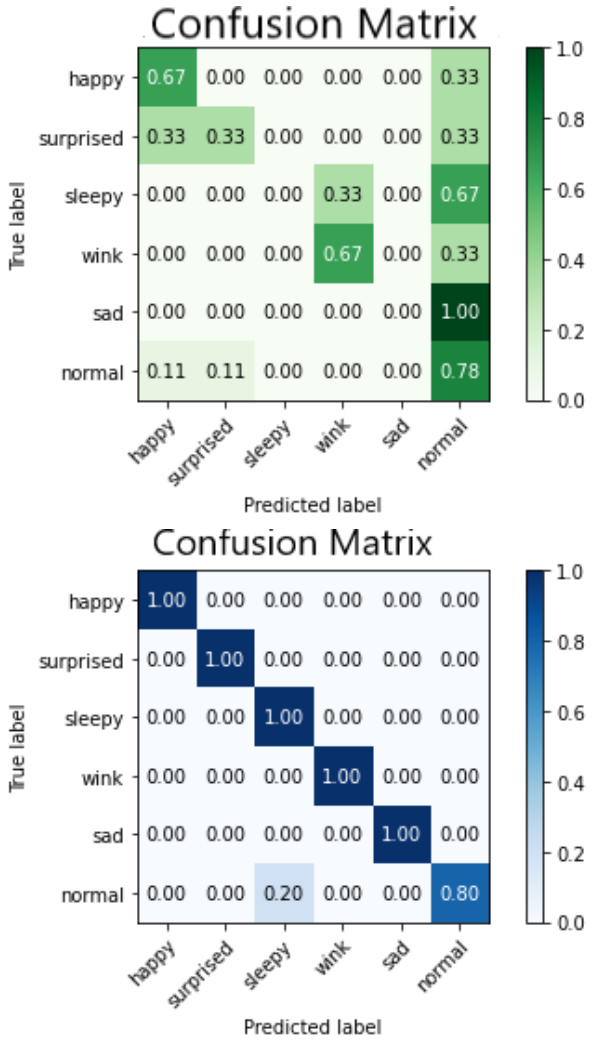

Figure 4.21 The normalized confusion matrix plots of the CNN (up, green) and the proposed hybrid FER model (down, blue) on Yale-Face test data

the early stop can cause the model to oscillate in training. To avoid temporary instability assigning a low patience level can be effective since the model can stop early with a good training accuracy without over-fitting.

The KDEF dataset yielded 692 images reflecting seven distinct facial emotions. Unlike the CK+ dataset, the KDEF dataset's images are not FACS encoded. The distribution of samples in each emotion class is shown in Fig. 4.10. The CK+ dataset is trained using the same method. The CK+ dataset provided 423 images of seven different facial expressions. The distribution of samples in each emotion class is shown in Fig. 4.9. Since the CK+ dataset is FACS coded, the chances of achieving a reasonable accuracy rate are high. For three datasets, the normalized confusion matrices are obtained. The

Table 4.6 Comparison of the Proposed model with other important existing models on Yale-Face dataset.

Author	Year	Methodology	Number of Classes/ No of Samples	Accuracy	Testing Procedure
Xie et al. [106]	2008	spatially maximum occurrence model (SMOM) and elastic shape–texture matching (ESTM) SMOM-ESTM	5 Classes (normal, happy,surprise, sleepy,wink)	94.7%	Leave-one-out mechanism
Xie et al. [106]	2008	Elastic shape–texture matching (ESTM)	5 Classes (normal, happy,surprise, sleepy,wink)	73.3%	Leave-one-out mechanism
Xie et al. [106]	2008	PCA	5 Classes (normal, happy,surprise, sleepy,wink)	65.3%	Leave-one-out mechanism
Xie et al. [106]	2008	SMOM	5 Classes (normal, happy,surprise, sleepy,wink)	92%	Leave-one-out mechanism
Platt et al. [107] Shan et al. [108] Lin et al. [27]	2009	LBP + DDAG TM Decision Directed Acyclic Graph	4 Classes (happiness, neutral, sadness, surprise)	56.95%	Ten-Fold-Cross-Validation
Friedman [109] Shan et al [108] Lin et al. [27]	2009	LBP + MaxWins TM	4 Classes (happiness, neutral, sadness, surprise)	57.26%	Ten-Fold-Cross-Validation
Platt et al.. [107] Lin et al. [27]	2009	LBP + DDAG SVM	4 Classes (happiness, neutral, sadness, surprise)	69.42%	Ten-Fold-Cross-Validation
Friedman [109] Lin et al. [27]	2009	LBP + MaxWins SVM	4 Classes (happiness, neutral, sadness, surprise)	68.42%	Ten-Fold-Cross-Validation
Platt et al. [107] Lin et al. [27]	2009	PCA + LBP + DDAG SVM	4 Classes (happiness, neutral, sadness, surprise)	62.31%	Ten-Fold-Cross-Validation
Platt et al.. [107] Lin et al. [27]	2009	2DPCA + DDAG SVM	4 Classes (happiness, neutral, sadness, surprise)	66.20%	Ten-Fold-Cross-Validation
Friedman [109] Lin et al. [27]	2009	2DPCA + MaxWins SVM	4 Classes (happiness, neutral, sadness, surprise)	65.60%	Ten-Fold-Cross-Validation
Lin et al. [27]	2009	Mixed-feature model (LBP+2DPCA projection features+DDAG-based SVM)	4 Classes (happiness, neutral, sadness, surprise)	81.28%	Ten-Fold-Cross-Validation
Hedge and Seetha [110]	2017	Subspace based FER using Combinational Gabor based Feature Fusion	6 Classes (happiness, neutral, sadness, surprise, wink,sleepy)	87.79%	Leave One Out Technique
Nigam et al. [111]	2018	HOG in Wavelet domain	4 Classes (happiness, neutral, sadness, surprise)	75%	Three fold leave k - samples out scheme (k=5)
Ravi et al. [114]	2020	CNN	All the samples	31.82%	Hold-Out Validation (70%-Training 30%-Test)
Our Proposed model	2021	Multi-Feature Fusion Deep Neural Network (HOG+VGG-Face)	6 Classes (happiness,surprise, sadness,wink,sleepy, normal)	95.26%, 96.67%	Five and Ten Fold Cross-Validation

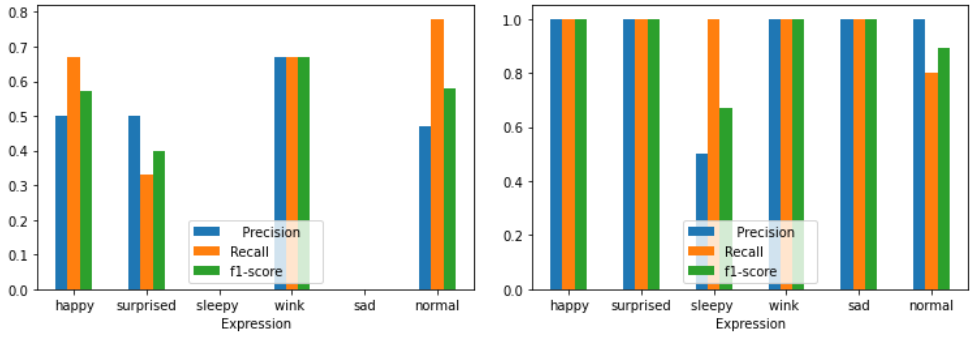

Figure 4.22 Comparison of Precision, Recall, and F1- score information of the CNN model (left) and Multi-Feature fusion model (right) on Yale- Face data

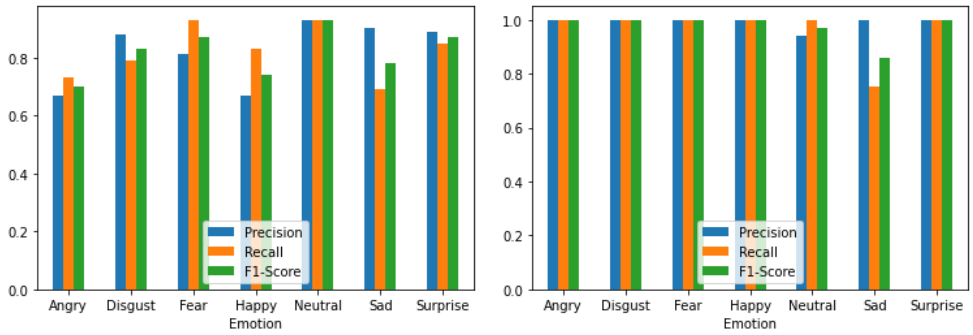

Figure 4.23 Comparison of Precision, Recall, and F1- score information of the CNN model (left) and Multi-Feature fusion model (right) on CK+ data

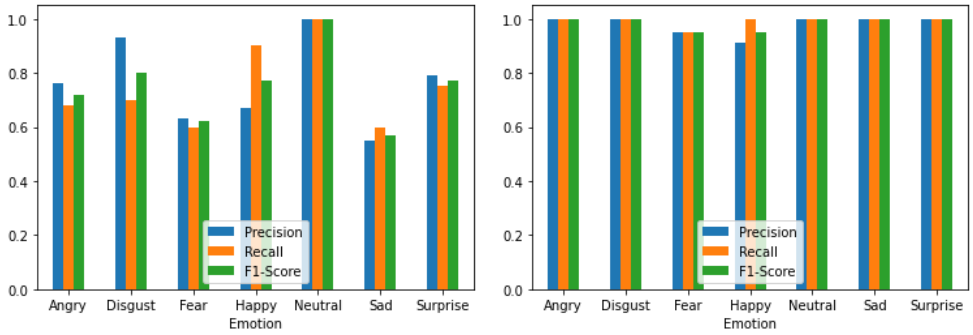

Figure 4.24 Comparison of Precision, Recall, and F1- score information of the CNN model (left) and Multi-Feature fusion model (right) on KDEF data

Table 4.7 Comparison of the Proposed model with other important existing models on CK+ dataset.

Author	**Year**	**Methodology**	**Number of classes**	**Number of samples**	**Accuracy**	**Testing technique**
Liu et al. [36]	2014	Facial Expression Recognition via Boosted Deep Belief Network	8	1308	96.7%	leave-one-subject-out training/testing strategy
Lv et al. [37]	2014	Face parsing using Deep Belief Neural network and Autoencoder	7	593	91.11%	7-fold cross-validation
Mollahosseini et al. [38]	2015	GoogLeNet and AlexNet inspired architectures	7	309	93.2%	5-fold-cross-validation
Lin et al. [112]	2015	A two-stage multitask sparse learning (MTSL) framework	6	96 subjects related sequences	93.46%	10-fold cross-validation
Khorrami et al. [39]	2017	Zero-bias CNN with FACS comparision	6	1308	95.1%	10-fold cross-validation
Zhang et al. [40]	2017	Hierarchical Bidirectional RNN	7	593	98.50%	10-fold cross-validation
Datta et al. [42]	2017	Hierarchical multi-class SVM architectures	7	593	91.85% (one vs one) 89.26% (DAGSVMs)	10-fold cross-validation
Nwosu et al [113]	2017	Two-channel convolutional neural network	7	350	95.72%	10-fold cross validation
Cai et al. [43]	2018	A novel Island loss using CNN and VGG16	7	981	94.39%	10-fold cross-validation
Xie et al [40]	2018	Deep comprehensive multi-patch aggregation CNN	6	927	93.46%	10-fold cross-validation
Kurup et al. [41]	2019	A semi-supervised emotion recognition algorithm with reduced features	7	1400	98.57% (HOG- Mouth)	10-fold cross-validation
Ravi et al. [114]	2020	Local binary patterns (LBP+SVM) and CNN	7	981	89.62%	hold-out validation 70% training 30% testing
Our Proposed model	2021	Multi-Feature Fusion Deep Neural Network (HOG+VGG-Face)	7	423	98.12%, 98.11%	5,10-fold cross-validation

Table 4.8 Comparison of the Proposed model with other important existing models on KDEF dataset.

Author	**Year**	**Methodology**	**Number of classes**	**Information about pose**	**KDEF Accuracy**	**Testing method**
Alshamsi et al. [115]	2017	Facial Landmarks descriptor and the Center of Gravity descriptor.	7	Frontal poses	90.8%	Hold-out validation 70% traing and 30% testing
Koujan et al. [116]	2020	Novel 3D Morphable Model	7	Frontal poses	92.24%	5-fold cross validdation
Melaugh et al. [117]	2019	Convolutional Neural network	7	Frontal poses	89.4%	Hold-out validation
Our proposed model	2021	Multi-Feature Fusion Deep Neural network (HOG+VGG-Face)	7	Frontal poses	96.36%,97.84%	5,10-fold cross validdation

proposed hybrid FER model has achieved higher accuracy scores than the CNN model on three datasets. The accuracy score of the CNN model on the KDEF dataset is 75.84%, whereas the accuracy score of the proposed hybrid FER model is 96.36% (5-fold CV), refer to Fig. 4.20. The accuracy score of the CNN model on the CK+ dataset using five-fold cross-validation is 81.11%, whereas the accuracy score of the proposed hybrid FER model is 98.12%, refer to Fig. 4.19. The training of the model using both handcrafted and self-learned features has dramatically improved accuracy scores on CK+ and KDEF datasets. The Yale-Face dataset has provided 120 images of six classes. The yale face dataset, unlike CK+, has images of different lighting conditions. The dataset is challenging to train as it provides only 120 images. The proposed model is instrumental in training datasets with fewer samples since the model fuses two different unique features. The accuracy scores have increased significantly due to the increase in features. The training graphs of the two models are shown in Fig. 4.14. The CNN model has shown a poor classification rate on classes like sleepy and sad in the yale face dataset. The confusion matrix, refer to Fig. 4.21, during the cross-validation suggests the inefficiency of the CNN model to distinguish the expressions sleepy and normal. The expressions like wink and sleepy were also frequently incorrectly classified by the CNN model. It is observed that the proposed model has shown great improvement in classifying the expressions like sleepy and wink. The results of the experiment, refer to Figs. 4.21, 4.22 has shown that the proposed model outperforms the conventional models in FER on Yale-face dataset. The detailed classification accuracy rates of three models on three different datasets using 5-fold and 10-fold cross-validation techniques are shown in Table 4.5.

4.5.4 Comparison of the proposed model with other popular FER models

Table 4.5 explains the mean accuracy scores and standard deviation of classification accuracy scores using five-fold and ten-fold cross-validation of three models on CK+, Yale-Face, and KDEF datasets, and there is a significant increase in the accuracy score. The proposed model outperforms a conventional CNN model in terms of accuracy and other classification metrics, refer to Figs. 4.22, 4.23 and 4.24. The proposed model is compared with other existing FER models on three datasets, refer to Tables 4.6, 4.7 and 4.8. The important aspects like the number of samples, number of classes, methodology, and

Figure 4.25 Prediction variations of two models on KDEF test images

testing/validation procedure are also discussed in the comparison table. The combination of HOG and VGG-Face CNN features based FER model designed in this experiment achieved a better accuracy score on CK+, Yale-face, and KDEF datasets when compared with other important FER models discussed in Tables 4.6, 4.7 and 4.8. The prediction of emotions of the two models on KDEF test images is shown in Fig.4.25. The accuracy of the proposed model is seen in distinguishing the micro-expressions of KDEF data, refer to Fig. 4.25. The emotions of fear and surprise have some correlation because of the similar micro-expressions like an eyebrow-raiser. The emotions of sadness and anger also share similar facial expressions. In Fig.4.25, the proposed hybrid model, unlike the CNN model, successfully classified emotions by analyzing complex facial features.

4.6 Conclusion

The proposed FER model extracts crucial patterns of facial expressions using a combination of HOG and CNN features. Compared to a conventional convolutional neural

network, the proposed FER model has a better discriminating ability in classifying similar emotions. In a traditional convolutional neural network, emotions like anger are strongly correlated with other emotions like neutral and sadness. The faces displaying emotions like surprise are also often misclassified as fear by the CNN model. The proposed model succeeds in classifying similar emotions, which is a main drawback of the CNN model. The classification metrics like precision, recall, and f1-score convey that the proposed model has significantly improved discriminating expressions of sleepy and sadness on the Yale-Face dataset. The comparison of classification metrics and normalized confusion matrices show that the proposed FER model outperforms the other existing models in classifying facial expressions on the CK+, Yale-Face, and KDEF facial expression datasets. The proposed FER model has achieved accuracy scores of 98.11%, 97.84%, and 96.67% on CK+, KDEF, and Yale-face facial expression databases using the 10-fold-cross-validation process.

Chapter 5

SternNet: A Rank of Confidence based Multi-Stage Facial Expression Classification Model

5.1 Introduction

The proposed SternNet is a multistage FER model which uses the rank of confidence (RoC) in predicting emotions. The deep convolutional neural networks face issues in getting trained on sparse and class imbalanced samples due to overfitting. Therefore we are not sure how credible the CNNs are during the prediction process. The confidence score cannot be a metric in deciding the credibility because overfitted DNN models usually express wrong confidence scores in predicting emotions. Therefore we need to have other metrics to judge how confident is the prediction given by the DNN model. The proposed model Sternnet imparts stern rules in classifying emotions. The details of the RoC and multistage SternNet model are described in this section. Initially, we perform different experiments on CNN model in training the sparse and class imbalanced samples. The AUC Area under curve of different classes is examined on conventional CNN models. The later section of the work discusses the drawbacks of conventional CNN models in predicting emotions. This work exclusively uses subject-independent CK+ data of six emotion classes for all the experiments.

Types of FER models Automatic FER models are first designed using handcrafted features. Handcrafted features depend on important corners, edges, and other salient

geometrical patterns present on the face. The examples of handcrafted features are local binary pattern [118], harris corner points [119] and HOG features [120]. The FER models initially extract the handcrafted features and are given to a machine learning classifier to segregate the emotions according to the patterns observed in the expressions. These conventional FER models depend on the geometry and are usually biased to the training data. At present, the advanced FER models use deep learning methods to improve the model's accuracy and get an unbiased method of analyzing emotions. Deep neural networks like convolutional neural networks are used in designing advanced FER models. These state-of-the-art FER models extract important patterns observed in different facial expressions directly from the dataset. The series of convolutional layers extract low and high-level features from the images and try to classify emotions even in noise. This adaptable nature of convolutional neural networks made it a more robust and accurate classifier. The deep learning-based FER models require a massive dataset for proper training of the neural network. It also requires a considerable amount of time to train the neural network and quick training requires a sophisticated GPU for its huge computation process. The deep neural networks are prone to misclassification if the dataset possesses a class imbalance. By constructing artificial datasets with different combinations of complexity, training set size, and degrees of imbalance, Japkowicz [121] investigated the effects of class imbalance. In the 1990s, Anand et al. [122] investigated the impact of imbalanced datasets on the backpropagation algorithm in shallow neural networks. The authors show that in class imbalanced situations, the gradient component of the minority class is much smaller than the gradient component of the majority class. In other words, the majority class controls the net gradient, which is in charge of updating the model's weights. During early iterations, this decreases the error of the majority group easily, but it also raises the error of the minority group, causing the network to become trapped in a slow convergence mode.

Problem Statement In this chapter, the authors discuss the crucial drawbacks of the conventional convolutional neural networks in classifying facial expressions. The classification of emotions is tricky since it involves distinguishing different emotion classes that share similar facial expressions (facial muscle movements). The difficulty of classifying emotions on different faces varies depending on the subject. Few subjects tend to express emotions that are easy to classify and others its tough to classify. The SternNet

model designed in this work tries to understand whether the sample is easy or difficult to classify using an accuracy measuring parameter known as "Rank of Confidence". The Rank of Confidence analyzes the face spatially in different views and scores a value that can express the confidence of the prediction. The high confident samples are easy to be classified correctly and low confident samples are difficult to be classified correctly. The SternNet stage-1 deals with the high confident samples and the low confident samples need further advanced evaluation for correct prediction hence it is dealt with advanced FER models in SternNet stage 2.

The work done in this chapter initially implements a conventional CNN model and tries to evaluate the performance of the classification on a class imbalanced sparse facial expression dataset. The drawbacks of the CNN model are analyzed and the steps taken for improving/ modifying the CNN model is discussed in this chapter. The proposed design is then compared with CNN model and other popular FER models to evaluate the efficiency of the designed model.

Organisation of the Chapter Section 5.2 implements a conventional CNN model (VGG16) in classifying facial expressions on extended CohnKanade facial expression database (CK+). Section 5.2.2 discusses the reasons for the low accuracy of CNN models in dealing with facial expressions. Section 5.3 demonstrates the design and advantage of SternNet model. The section focuses on explaining the crucial advantages of the rank of confidence. Section 5.4 discusses the results of different classifiers used in SternNet model and the comparison of the proposed model with other popular models is also explained in the section.

5.2 Experimenting with the convolutional neural network on subject independent, class imbalanced, and sparse CK+ dataset

The database for emotion classification is taken from extended Cohn-Kanade(CK+) [77] facial expression database. The extended Cohn-Kanade, widely known as CK+, is a

facial expression dataset for the classification of facial action units and is popularly used for facial emotion recognition. The dataset has posed as well as non-posed expressions. The Extended Cohn Kanade (CK+) dataset consists of 593 sequences across 123 different subjects. Each sequence in the database contains frames varying from 10 to 60, and in every sequence, the frames are captured such a way that there is a shift in expression from a neutral to the peak intensity of a specific emotion. Among the given sequences, only 327 sequences with 118 subjects have the expression labels of anger, contempt, disgust, fear, happiness, sadness, and surprise. In this experiment, we design the FER model to classify six emotions anger, disgust, fear, happiness, sadness, and surprise. Many papers include multiple frames of a sequence (last 5 or 3 frames) into the training and testing dataset. In this work, only the last image of the labeled sequence, which has the peak intensity of emotion, is chosen and taken into the dataset.

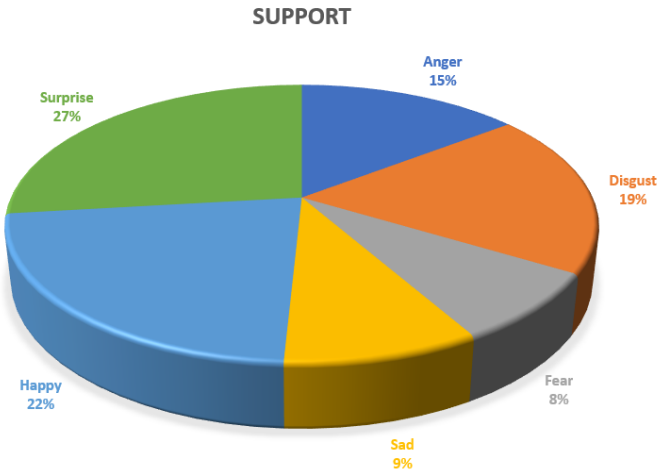

Figure 5.1 Information about the percentage of the number of samples in each class

We have used the transfer learning technique and imported the VGG16 [13] model. The imported model is loaded with weights that are pre-trained on the imagenet [61] dataset. The ending layers of the VGG16 are truncated, and new layers are attached to the network to match the output classes. To the base network, we have added a flatten layer, a dense layer of 256 neurons, and a dropout layer. It is followed by another dense layer with 128 neurons, a dropout layer, and a softmax layer of six output classes, refer to the model design in Fig. 5.17. The softmax layer is modified, and a softmax layer of six classes is added at the end of the CNN. The block diagram of the model is shown in Fig. 5.17. All the other layers except the added layers are frozen to use the pre-trained

Table 5.1 Information of the number of samples in each emotion class

Emotions Total No of Samples =309	**Number of Samples**	**Percentage of samples**
Anger	45	15%
Disgust	59	19%
Fear	25	8%
Happy	69	22%
Sad	28	9%
Surprise	83	27%

weights. Let y_i be the softmax score of i^{th} class of a sample then the equation 5.1 gives the softmax probability of prediction.

$$\textit{Softmax score for each class i= 1 to 6}: f(y)_i = \frac{e^{y_i}}{\sum_j^{N=7} e^{y_j}} \tag{5.1}$$

The information of samples in each class (support) is shown in Fig. 5.1. The samples that display the emotion of fear are only 8% in the dataset. The samples of fear and sadness are very less when compared to the samples of happiness and surprise. In this first experiment, we have used the VGG16 model to classify facial expressions on the extracted CK+ data. A five -fold-cross-validation method is used to analyze the classification accuracy in each emotion class. The normalized confusion matrix along with the Receiver Operator Characteristic (ROC) curves after the five-fold-cross-validation are shown in Fig.5.2. The precision, recall, and F1-score results are also calculated. The accuracy score obtained after five-fold-cross-validation is 81.57% (+/-2.10%). The results shown in the table 5.2 indicate that the classification of emotions such as fear and sad have a low Area Under the Curve (AUC) when compared to the AUC of happiness and surprise. The AUC Area Under the Curve (AUC) evaluates the ability of a model to distinguish between classes (emotions). The low AUC of emotions fear and sad are seen during the cross-validation in test-folds 1, 2, 4, and 5, refer Fig. 5.2.The class imbalance is one of the reasons for low accuracy rates seen in emotions fear and sadness. The normalized confusion matrices also indicate that the emotion sad is often predicted as anger

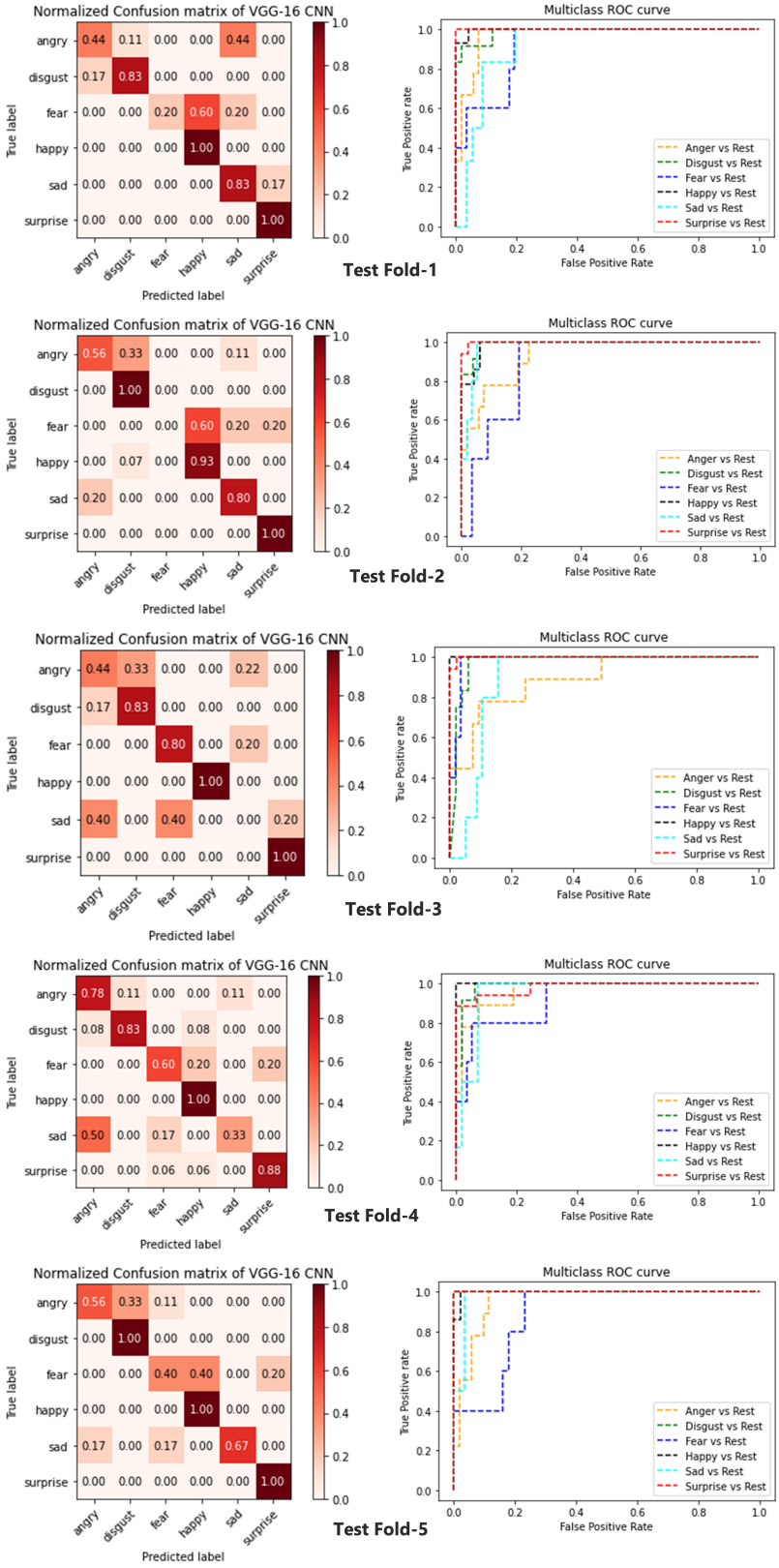

Figure 5.2 The obtained normalized confusion matrices and ROC curves of five test folds using VGG16

Table 5.2 Five-fold-cross-validation classification metrics of VGG-16 on subject independent CK+ test data

CK+ dataset 309 samples	**Precision**	**Recall**	**F1-Score**	**Test samples**	**AUC**
Anger	0.694	0.556	0.608	9	0.94
Disgust	0.826	0.898	0.856	11	0.99
Fear	0.554	0.4	0.42	5	0.92
Happy	0.864	0.986	0.92	14	1
Sad	0.568	0.526	0.518	6	0.95
Surprise	0.94	0.976	0.958	16	1

and vice-versa. In this experiment can be considered that the conventional CNNs tend to perform poorly at classifying facial expressions with sparse and class imbalanced samples.

To overcome these issues we have designed a novel FER model in this work. The designed model first tries to analyze the confidence of the prediction which means answering the question of how accurate is the prediction done by the FER model. On analyzing the confidence of prediction the model tries to decide whether the sample needs further evaluation. The designed model is named "SternNet". The model considers stern rules in classifying emotions and considers Facial Action Coding System (FACS) as an important topic.

5.2.1 Facial Action Coding System

The Facial Action Coding System (FACS) was created by Paul Ekman and Wallace V. Friesen to distinguish various facial expressions on any human face. FACS is used to deconstruct and properly taxonomize essential facial features based on their properties. FACS assisted in the development and description of "Action Units," which are discrete acts of muscles/muscle contraction and relaxation (AUs). As seen in table 5.3, six combinations of different action units on the face construct six different emotions.

Table 5.3 FACS information about action units for six emotions. [7]

Emotion	Facial Muscle	Corresponding Action Units
Anger	Brow lowerer+ Upper lid raiser+ Lid tightener+ Lip tightener	4+5+7+23
Disgust	Nose wrinkle+ Lip corner depressor+ Lower lip depressor	9+15+16
Fear	Inner brow raiser+ Outer brow raiser+ Brow lowerer+ Upper lid raiser+ Lid tightener+ Lip stretcher+ Jaw drop	1+2+4+5+7+20+26
Happiness	Cheek raiser+Lip corner puller	6+12
Sadness	Inner brow raiser+Brow lowerer+Lip Corner depressor	1+4+15
Surprise	Inner brow raiser+Outer brow raiser+ Upper lid raiser(Slight)+ Jaw drop	1+2+5B+26

Table 5.4 Intensity level variations in FACS. [7]

Alphabet	A	B	C	D	E
Intensity Level	Trace	Slight	Pronounced	Extreme	Maximum

5.2.2 Correlation between action units in different emotions

Facial Emotions are tough to classify since the problem is a subclassification task that involves identifying the emotional classes with a very slight difference. FACS has also measured the action unit's intensity by scaling the intensity levels with A to E letters, where A is the weakest and E is maximum intensity. In the FACS coding, refer to the table 5.3, it is evident that various emotional states have the same facial muscle moments, for example, disgust and sad emotions trigger the same Lip Corner Depressor (Action unit-16). There is a high probability of misclassifying the emotions due to these similarities in the different emotion classes. There is much difference in the emotions of happiness and surprise because there is no intersection of action units in both emotions. FACS can convey important information about the probability of accurately differentiating two emotions through the study of their respective action units. The inclusion of FACS information in the FER model improved the model's accuracy and helped in a better understanding of action units in emotion classification.

Table 5.5 One hot encoding of important action units that represent six emotions.

Action Units	Angry	Disgust	Fear	Happy	Sad	Surprise
AU1	0	0	1	0	1	1
AU2	0	0	1	0	0	1
AU4	1	0	1	0	1	0
AU5	1	0	1	0	0	1
AU6	0	0	0	1	0	0
AU7	0	0	1	0	0	0
AU9	0	1	0	0	0	0
AU12	0	0	0	1	0	0
AU15	0	1	0	0	1	0
AU16	0	1	0	0	0	0
AU20	0	0	1	0	0	0
AU23	1	0	0	0	0	0
AU26	0	0	1	0	0	1
AU27	1	0	0	0	0	0

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

On analyzing the action units in emotions, the majority of them lie on the important facial landmarks like eyebrows AU(1, 2, 4), eyes AU(5,7), and lips AU(23, 16, 26, 12, 23), refer to table 5.3. On performing exploratory data analysis of the FACS action units, the correlation of action units on upper and lower facial parts is calculated. The exploratory data analysis done on action units in this work has evaluated the correlation matrix of different emotions on upper and lower facial parts using the Pearson correlation coefficient. The upper face consists of eyes, nose, and cheeks as important facial landmarks and the lower face has a mouth and some part of the cheeks as an important landmark. The spatial correlation analysis of action units is shown in the form of heatmap in the Figs 5.4, 5.3. On analyzing the Pearson correlation matrices the emotions sad and disgust have a good correlation at the lower face since they share similar action units, but they have shown poor correlation at the upper face since disgust has a unique action unit- 9 on the nose. So the emotion of disgust shows its dominant features on the upper face when compared

to the lower face. The correlation coefficient has a value between -1.0 and 1.0. The power of association is determined by the correlation coefficient's value. If the correlation coefficient is between 0.5 and 1.0, it means there is a good positive relationship. A small association is indicated by a correlation coefficient between 0 and 0.5. A negative correlation is implied by a correlation coefficient between 0 and -1.0. The emotions of fear and surprise share similar action units and hence there may be a high chance of misidentifying the emotions of fear as a surprise and vice-versa. These correlations seen in different emotions make the task of classifying emotions hard. So, we need to design a special model that analyzes action units in different facial regions. In the figures 5.4 and 5.3 the correlation of emotions on upper and lower face regions are displayed in the form of correlation heat-map. The action unit and emotion heat-maps of lower and upper face regions clearly show the correlation of different emotions, for example on the upper face there exists correlation between fear and surprise. In the lower face heat-map the emotions of sad and disgust also show a high correlation. The existence of this correlation makes the classification of micro-expressions difficult. In this next section, we explain different methods to tackle these issues using SternNet model.

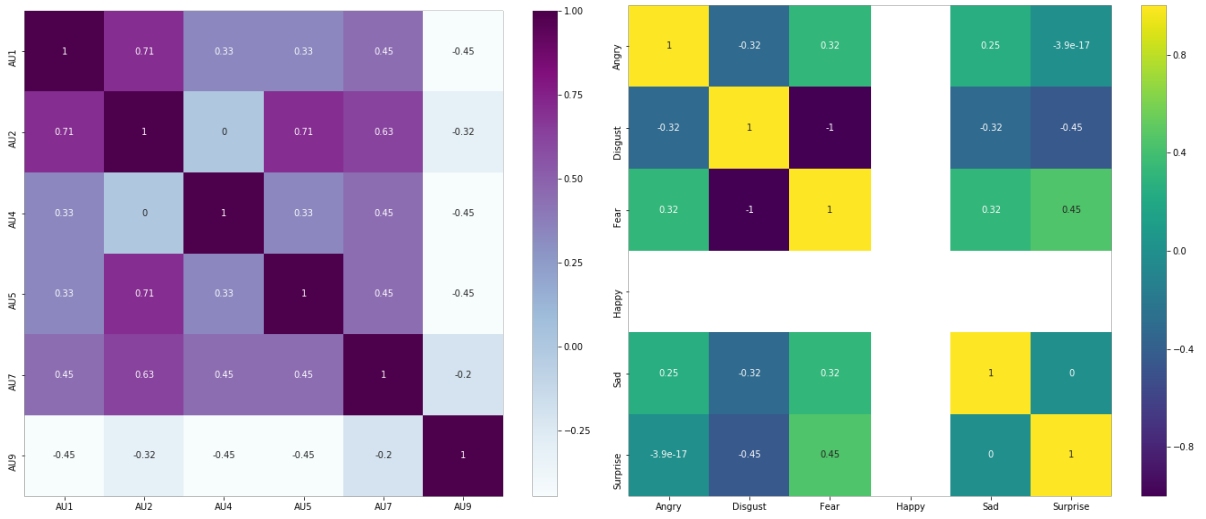

Figure 5.3 Correlation of different action units and emotions on upper face

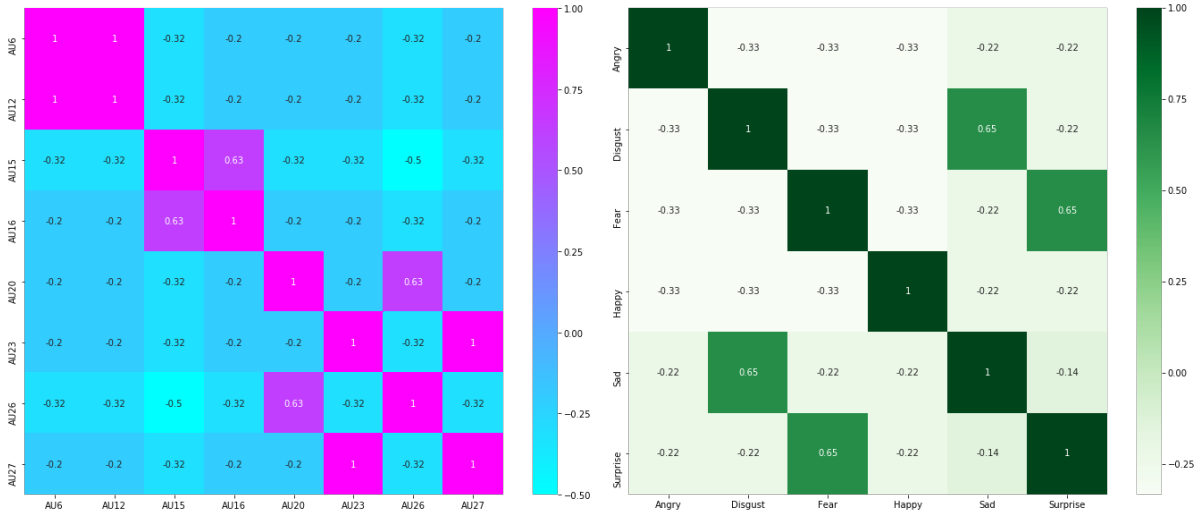

Figure 5.4 Correlation of different action units and emotions on lower face

Figure 5.5 Partition of face

5.3 SternNet: A multi-stage FER model

The objective of the SternNet is to separate confident samples and unconfident samples. The sample is termed as confident if it posses a high probability of being correctly classified. In the same way, a sample is defined as unconfident if the sample is prone to misclassification. The SternNet analyzes different facial regions before it decides the confidence. The face can be partitioned in various ways. The face partition is done in three ways. The partition is done in such a way that the obtained spatial regions contain different combinations of action units. The Fig. 5.6 shows the distribution of action units spatially on a face.

Various data pre-processing steps were taken before applying to the SternNet. The pre-processing of the images is discussed in the below points.

1. Face Detection: The pixels in the images should contain only the face. To extract

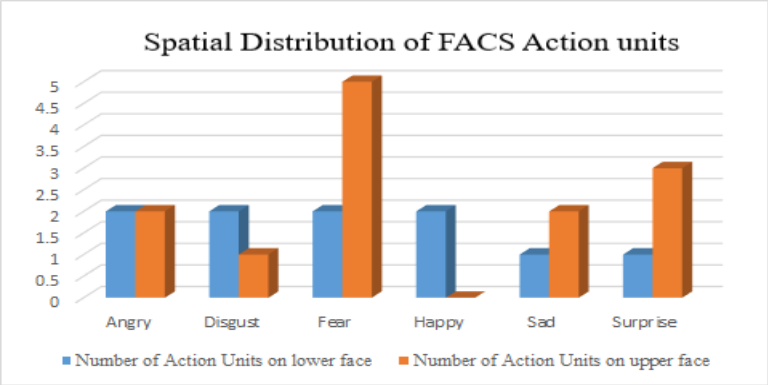

Figure 5.6 Distribution of action units on upper and lower face

only facial region Viola-Jones [123] algorithm is used to extract the pixels that exclusively contain the face.

2. Image Normalization: The pixels are normalized before applying to the classifier. The range of pixels is converted from (0 to 255), to (0 to 1) by dividing each pixel value by 255 and the image size is resized to (224,224)

3. Face Partition: To obtain three different facial regions, we have used dlib's facial landmark [66] library to extract the region of interests (RoIs), refer Fig.5.5.

4. The first facial region known as "Lower face-1" contains only the mouth as an important landmark, refer Fig. 5.10. The action units observed in this facial region are AU 10- Upper Lip Raiser, AU 12- Lip Corner Puller, AU 15- Lip Corner Depressor, AU 17- Chin Raiser, AU 20- Lip Stretcher, AU 23- Lip Tightener, AU 24- Lip Pressor, AU 26- Jaw drop, and AU 27- Mouth Stretch.

5. The second facial region known as "Upper face" contains eyes, eyebrows, and nose (refer Fig. 5.10). The action units observed in this region are AU 1- Inner Brow Raiser, AU 2- Outer Brow Raiser, AU 4- Brow Lowerer, AU 5- Upper Lid Raiser, AU 7- Lid Tightener, and AU 9- Nose Wrinkler.

6. The third facial region known as "Lower face-2" contains mouth, nose, and cheeks as important landmark points(refer Fig. 5.10). The important action units seen in addition to action units related to mouth are AU 6- Cheek Raiser and AU 9- Nose Wrinkler.

Figure 5.7 Lower face-1 region (important landmark mouth)

Figure 5.8 Upper face region (important landmarks eyes, eyebrows, and nose).

Figure 5.9 lower face-2 region important landmarks nose, cheeks and mouth.

Figure 5.10 Different facial regions considered by StrenNet

5.3.1 SternNet Stage-1 : Finding the Rank of Confidence

To increase the samples in a sparse dataset, we have created four different images for each sample according to the distribution of action-units . Every face has now four unique spatial information i.e, full face, lower face-1, upper face, and lower face-2. In the first stage, we extract the Histogram of Oriented Gradients features [124] of four images for each sample. The HOG features visualization in four facial regions full face, lower face-1, upper face, and lower face-2 are shown in the Figs. 5.11, 5.14 We have trained four SVM classifiers to learn patterns using HOG features. All the SVM classifiers use the linear kernel and implement squared hinge as the loss function. For the HOG features,

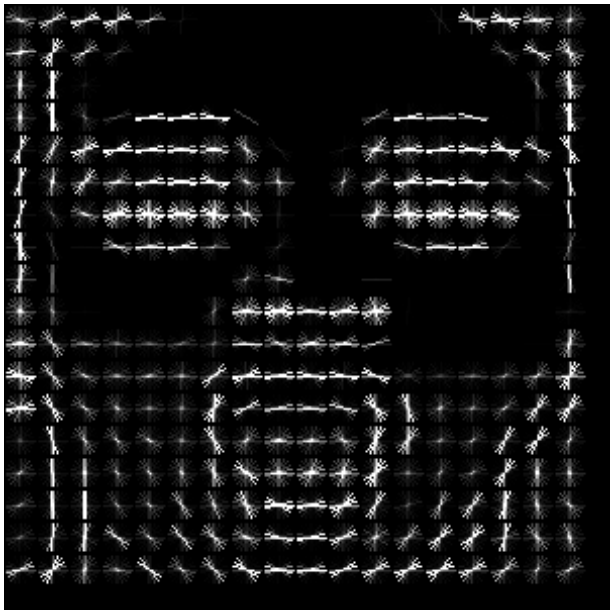

Figure 5.11 HOG visualization of full face region

the orientations are taken as 11, pixels per cell as (16,16) and cells per block are taken as (2,2). The SVM classifiers are trained with 247 images and tested with 62 images. The same process is repeated and we have implemented a 5-fold-cross-validation in all the four SVM classifiers. Now using the results of the four SVM classifiers we have allotted "Rank of Confidence" (RoC). The SternNet aims to segregate high and low-confident samples. Each sample is viewed from four different perspectives. A sample classified as emotion-1 by a support vector machine(SVM) classifier-1 may or may not be classified as emotion-1 by the other SVM classifiers. The sample is given the rank of confidence as 1 if all the four SVM classifiers predict the same emotion class. It means that the prediction given by the four SVM samples which are trained on four different HOG features (different combinations of action units) match without any discordance. In the same way, a sample is given a rank of confidence as 2 if any three SVM classifiers predict the same emotion class. The samples having a rank of confidence 2, 3 are prone to misclassification since there exists discordance between the classifiers. The pseudo-code is explained below, the predicted emotion class from each SVM classifier is read and according to the accordance, the rank of confidence (RoC) is assigned to each sample. Nearly average 73% of samples in the test data possess a rank of confidence of 1, which means 73% of total samples are high confident samples in the SternNet stage-1. The advantages of knowing the RoC can be useful in easy evaluation and segregation of samples according to the confidence.

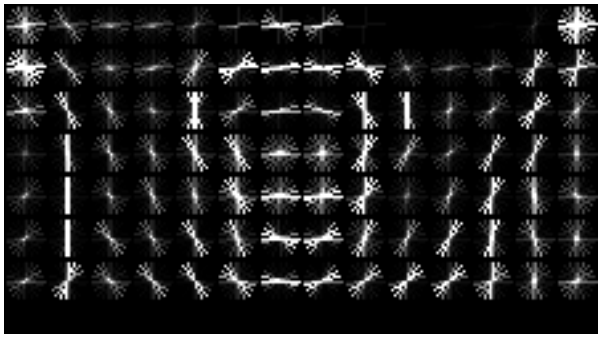

Figure 5.12 HOG visualization of Lower face-1

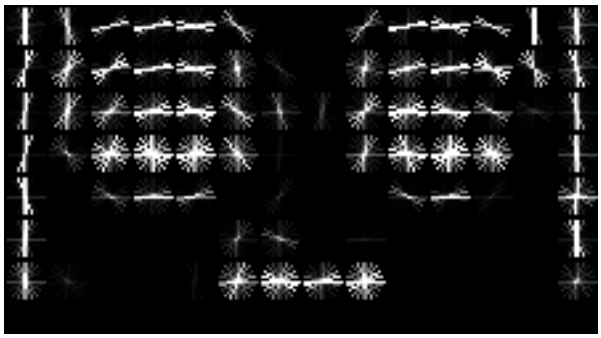

Figure 5.13 HOG visualization of Upper face

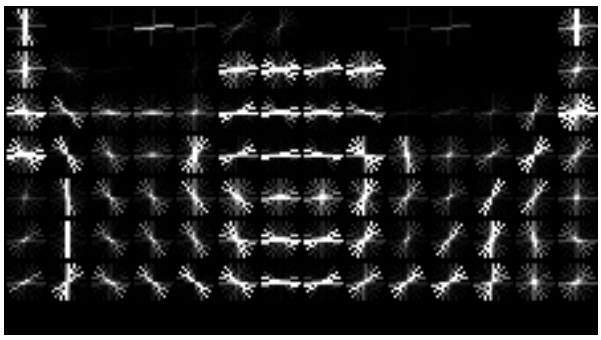

Figure 5.14 HOG visualization of Lower face-2

After completing the SternNet stage-1, we obtain the RoC of each sample. In StrenNet, if we consider extreme rules then only RoC-1 samples are considered as confident samples. But it is noted that always RoC-1 samples have a higher probability of correct prediction when compared to RoC-2 samples. Let us analyze the experimental results after SternNet stage-1. The total number of samples present in testing data in each fold is 62. The experimental results after five-fold cross-validation have shown that the predicted samples that possess RoC-1 are nearly 99% correct, refer to table 5.6. In test folds 1 and 2 the samples predicted to have RoC-1 are 100% correct. The important thing to be noted here is the samples exclusively obtained from RoC-1 have been predicted at an accuracy of nearly 99% (refer Table 5.6) whereas the accuracy using a single full face-

Table 5.6 SternNet stage-1 Analysis

5-Fold-Cross-Validation	Test Fold 1	Test Fold 2	Test Fold 3	Test Fold 4	Test Fold 5
Total number of test samples	62	62	62	62	61
Number of sample having Rank of Confidence =1	44	42	44	46	51
Correctly classifed RoC =1 samples	44	42	43	45	50
Accuracy of Prediction	100%	100%	97.27%	97.82%	98.04%

SVM classifier achieved an accuracy of 89% (refer Fig. 5.23. After completion of the five-fold-cross-validation of SternNet stage-1, on average, 73% of samples that are confident are predicted and the remaining samples which are unconfident are further evaluated in SternNet stage-2.

5.3.2 SternNet Stage 2 : Dealing with low Confidence samples

The discordance between the predictions of different SVM classifiers results in low confidence in the prediction and thus these low confidence samples are highly prone to misclassification. The HOG features derived from these samples are not sufficient to analyze the patterns and therefore other features are necessary to understand the micro-expressions. In stage - 2 of SternNet in addition to the handcrafted HOG features, self-learnt features from CNN are considered. We have designed a multi-feature fusion deep neural network model that considers HOG and VGG-Face [125] features for predicting emotions. The output emotion class of all the low confident samples is predicted by the multi-feature fusion network. In SternNet stage-2, in addition to the obtained HOG features of full-face region, the VGG-Face (refer Fig. 5.17) features are extracted from the training and testing samples. We have loaded VGG-Face pre-trained weights that are trained with massive facial image datasets like Internet Movie Data Base (IMDB) celebrity list, LFW, and YTF datasets. The images are resized to (224,224) to match the VGG-Face architecture. The HOG parameters such as the number of orientations of the gradient directions and cell size are changed in SternNet stage -2. The number

Algorithm 1 SternNet stage 1

procedure FINDING THE RANK OF CONFIDENCE OF TEST SAMPLES IN EACH TEST FOLD
    $f \leftarrow$ predicted emotion *by SVM-1 on test samples of full face*
    $l \leftarrow$ predicted emotion *by SVM-2 on test samples of lower face1*
    $m \leftarrow$ predicted emotion *by SVM-3 on test samples of lower face2*
    $u \leftarrow$ final emotion *by SVM-4 on test samples of upper face*
high confidence samples:
    for i, j, k, p in $zip(m, u, l, f)$ **do**
        if $i == j == k == p$ **then**
            $RoC = 1 \leftarrow sample$
            Rank of Confidence: *Label the samples as RoC-1*
    for i, j, k, p in $zip(m, u, l, f)$ **do**
        if $(i == j == k! = p) or (p == i == j! = k) or (p == i == k! = j) or (p == j == k! = i)$ **then**
            $RoC = 2 \leftarrow sample$
            Rank of Confidence: *Label the samples as RoC-2*
low confidence samples:
    Rank of Confidence: *The remaining samples are labeled as RoC-3*

Figure 5.15 Pseudo Code for the SternNet stage-1

Algorithm 2 SternNet stage 2

procedure DEALING WITH LOW CONFIDENCE SAMPLES
  $sl \leftarrow$ Self-learnt features *extracted from VGG-Face CNN network on test samples of full face*
  $hc \leftarrow$ Hand-crafted features *extracted from HOG features on test samples of full face*
high confidence samples:
  if $Rank\ of\ Confidence == 1$ **then**
    $SternNet\ stage-1 \leftarrow Testing\ the\ sample$
    $HOG\ features(hc) \leftarrow Extraction\ of\ features$
    $Four\ SVM\ classifiers \leftarrow Model\ used\ to\ train\ and\ test\ the\ samples$
    Confident Samples: *Label the samples*
Low confidence samples:
  if $Rank\ of\ Confidence > 1$ **then**
    $SternNet\ stage-2 \leftarrow Testing\ the\ sample$
    $VGG-Face\ and\ HOG features(sl\ and\ hc) \leftarrow Extraction\ of\ features$
    $A\ Multi-feature fusion\ DNN \leftarrow Model\ used\ to train\ and\ test\ the\ samples$
    Unconfident Samples: *Label the samples*

Figure 5.16 Pseudo Code for the SternNet stage-2

of orientations is taken as 9 in stage-2 and the cell size is taken as (8,8). The number of features is increased in SternNet stage-2 when compared to its stage-1. The Fig. 5.18 explains the StrenNet stage -2 procedure. The VGG-Face features of 2622 are extracted for the first input of the model and the second input, 15488 HOG features are extracted from the unconfident samples. The overall flowchart of the SternNet model is explained in Fig. 5.19 .

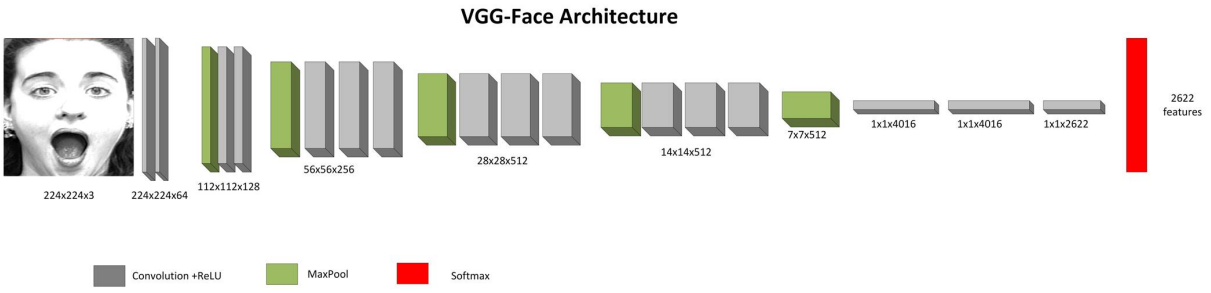

Figure 5.17 VGG-Face Architecture

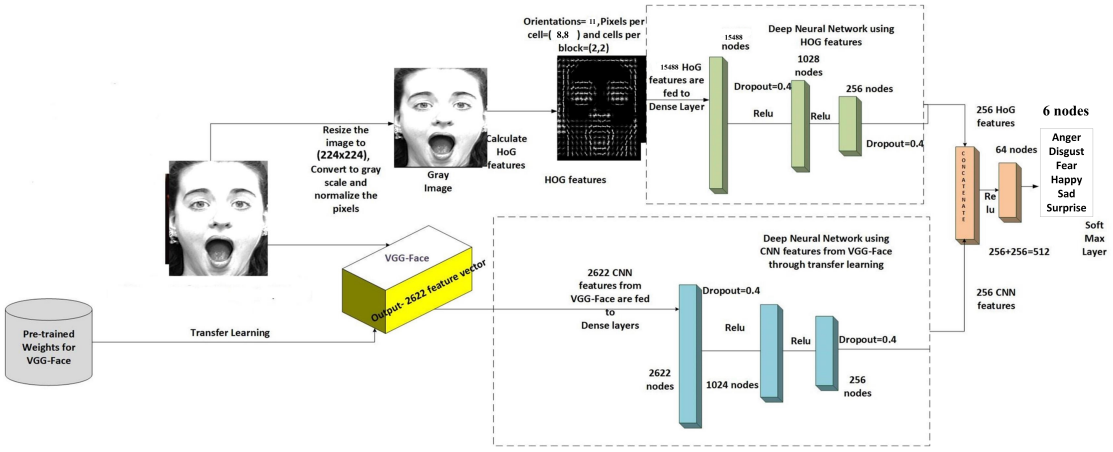

Figure 5.18 Multi-feature fusion model, combination of VGG-Face and HOG features at SternNet stage-2

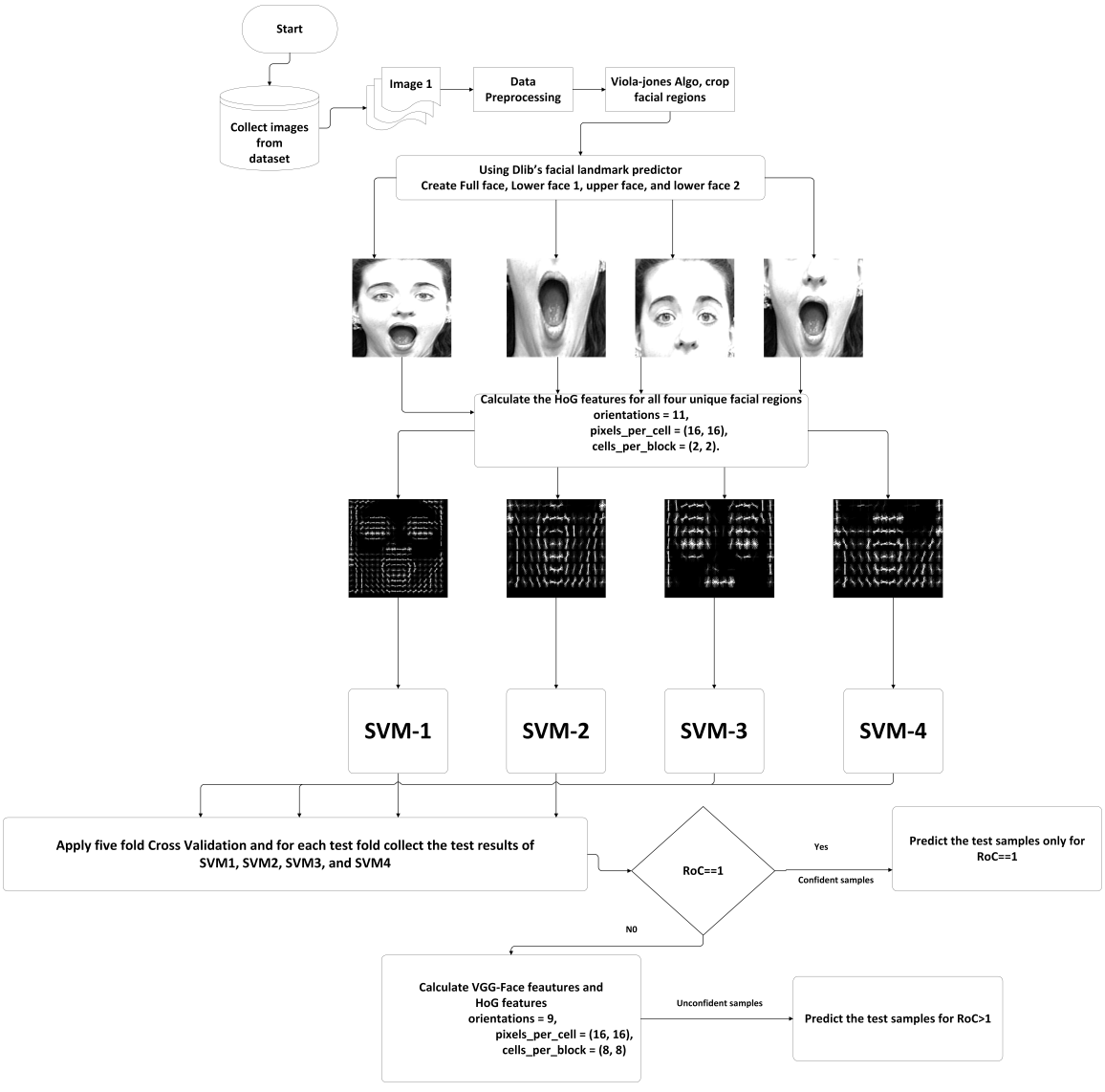

Figure 5.19 Flowchart of StrenNet model

5.4 Results and Discussion

5.4.1 Training and model validation

We have used Google's Colaboratory [126] GPU to train our model. Google's Colab provides GPU Nvidia 1xTesla K80, having 2496 CUDA cores and CPU Xeon Processor of the frequency of 2.3 GHz. Input images are resized to (224x224) as the VGG-Face model is trained using (224x224) sized images. The RMSprop optimizer is used in training the model. The equations used in RMSprop for updating the weights in the neural network are shown in equations 5.3,5.4. RMSprop helps to reduce unwanted oscillations and improves the speed of convergence. The important hyperparameters used in the RMSprop are decay rate (β), learning rate, and epsilon. The loss function categorical cross-entropy is used as an error function for training the weights of the neural layers. The cross-entropy loss function is widely used in error function in classification problems for deep neural networks [68]. If t_i and y_i be the target and the output predicted score of i^{th} class of a sample.

$$Categorical\ Cross\ entropy\ error: -\sum_{i=1}^{N=6} t_i log(y_i) \tag{5.2}$$

In Multi-Class classification problems, the targets are one-hot encoded, making only the positive emotion class appear in the categorical loss function.

$$E\left[g^2\right]_t = (\beta)E\left[g^2\right]_{t-1} + (1-\beta)g_t^2 \tag{5.3}$$

β- decay rate is taken as 0.9, g_t- Gradient at time t, $E\left[g^2\right]_t$- Exponential Average of squares of gradients.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E\left[g^2\right]_t + \epsilon}} g_t \tag{5.4}$$

ϵ- epsilon is taken as 10^-8, η- Learning rate is taken as 0.0001

5.4.2 Classification metrics

The important evaluation metrics of the FER model discussed are Accuracy, Precision, Recall, and F1-score. Let TP represents True Positives, FP represents False Positives, FN represents False Negatives, and FP represents False Positives.

1. **Accuracy**: Accuracy (Acc) is useful in evaluating model performance. However, when there exists a class imbalance problem, it is necessary to consider other important metrics like precision and recall.

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.5}$$

2. **Precision**: The precision (P) highlights the ability of the model to pick the desired class. P depends on TP and FP. False Positives are the number of predictions the model misclassifies as positive when the true label is negative.

$$P = \frac{TP}{TP + FP} \tag{5.6}$$

3. **Recall**: Recall (R) is the other classification metric that conveys the ability of the model to predict all the classes of interest in a dataset. R depends on TP and FN. FN is the number of predictions the model misclassifies as negative when the true label is positive.

$$R = \frac{TP}{TP + FN} \tag{5.7}$$

4. **F1 Score**: It is necessary to maintain good precision and recall for any model. The goal of a good classifier is to pick the correct class without any mistake (precision) and, at the same time, pick as many as correct classes (recall). A good trade-off is to be maintained between precision and recall. F1 score provides a decent blend of two metrics recall, and precision. F1 score is the harmonic mean of recall and precision.

$$F1\ Score = 2 * \frac{P * R}{P + R} \tag{5.8}$$

Table 5.7 SternNet accuracies at stage-1 and stage-2

SternNet stage	Model	Mean Five fold Cross Validation Accuracy
Stage-1	SVM Lower face-1 (HOG)	84% (+/- 2.13%)
	SVM Upper face (HOG)	81% (+/-1.18%)
	SVM Lower face-2 (HOG)	88% (+/- 2.26)
	SVM Full face (HOG)	89% (+/-2.42%)
	Mean Number of Samples Predicted in test Data (%) (HOG)	Accuracy
	(73%) 45 samples	98.62% (+/ 1.65%)
Stage-2 Overall Accuracy	VGG-Face+HOG Multi fusion DNN (100%) 62 samples	98.1%

Table 5.8 Classification Metrics of SternNet model after five-fold-cross-validation

Emotions	**Precision**	**Recall**	**F1-score**	**Support**
Anger	1	0.978	0.988	9
Disgust	1	1	1	12
Fear	1	0.92	0.938	5
Happy	0.986	0	0.992	13
Sad	0.966	0	0.982	6
Surprise	0.988	0.988	0.988	17

	precision	recall	f1-score	support
angry	0.88	0.78	0.82	9
disgust	0.69	0.75	0.72	12
fear	0.75	0.60	0.67	5
happy	0.86	0.86	0.86	14
sad	0.71	0.83	0.77	6
surprise	1.00	1.00	1.00	16
accuracy			0.84	62
macro avg	0.81	0.80	0.81	62
weighted avg	0.84	0.84	0.84	62

Figure 5.20 Classification metrics of SVM (lower face-1)

	precision	recall	f1-score	support
angry	0.64	0.78	0.70	9
disgust	1.00	0.92	0.96	12
fear	1.00	1.00	1.00	5
happy	0.93	1.00	0.97	14
sad	0.33	0.20	0.25	5
surprise	1.00	1.00	1.00	17
accuracy			0.89	62
macro avg	0.82	0.82	0.81	62
weighted avg	0.88	0.89	0.88	62

Figure 5.21 Classification metrics of SVM (lower face-2)

5.4.3 SternNet Model Analysis

The designed model has two stages, in the first stage, four SVMs are used to analyze emotions at different sections of the face that represent a unique combination of action units. The accuracy results show that lower face-2 is a very easy and useful spatial region on the face to understand emotions when compared to the upper face. The upper face (eyebrows, eyes, nose) can predict emotions at an accuracy of 81% whereas the lower face-2 (nose, cheeks, and mouth) can predict emotions at an accuracy of 88%. The lower face-2 has achieved higher accuracy than that of the lower face-1 (only mouth). Emotions such as disgust contain unique action unit-9, nose wrinkle which is not seen in any other emotions, therefore lower face-2 (includes the nose, and cheeks) has achieved good accuracy in predicting emotion disgust. StrenNet uses the important characteristics of all the SVMs to pick samples accurately. The stage-1 of StrenNet decides the confidence of prediction and defines the sample whether it is simple or complex to predict. The stage-1 completes its task by assigning rank of confidence and predicting only the confident samples (RoC=1 samples). The samples which are complex to predict (RoC greater than 1) are moved to stage-2 for next analysis. The stage-2 of SternNet uses additional features. In stage-2

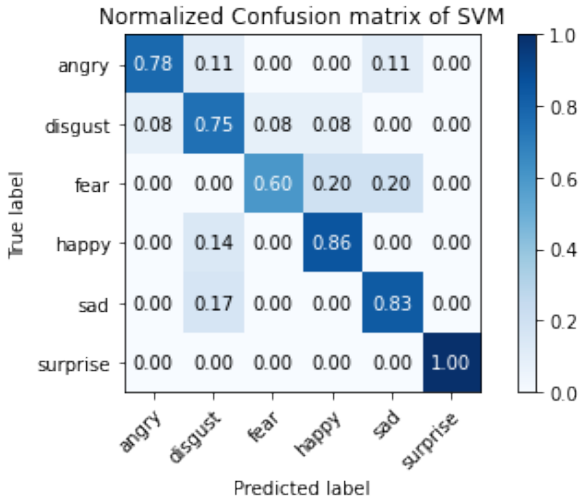

Figure 5.22 Normalized confusion matrix of SVM (lower face-1)

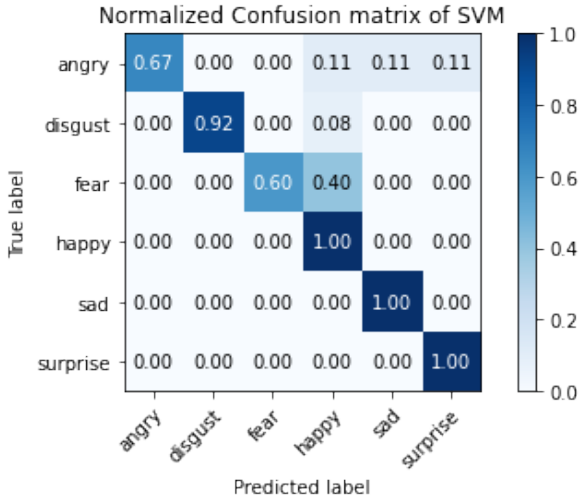

Figure 5.23 Normalized confusion matrix of SVM (lower face-2)

along with HOG features deep features are added to the feature collection for prediction. The stage-2 StrenNet uses different features when compared to stage-1 and uses deep neural network for emotion prediction. The Figs. 5.23, and 5.28 shows the normalized confusion matrices of individual SVMs used at the stage-1. The proposed model in this work explains the importance of using SternNet instead of using a single full-face SVM classifier. Stage-2 considers that the samples received are complicated to comprehend hence derive more features. The VGG-Face is an important model when compared to other models in the facial analysis since it is trained exclusively on faces of a massive dataset, whereas other models use the imagenet dataset. The deep neural network is built on a combination of VGG-Face and HOG features. The emotion class of all the

	precision	recall	f1-score	support
angry	0.67	0.89	0.76	9
disgust	1.00	0.92	0.96	12
fear	0.25	0.20	0.22	5
happy	0.93	1.00	0.96	13
sad	1.00	0.33	0.50	6
surprise	0.84	0.94	0.89	17
accuracy			0.82	62
macro avg	0.78	0.71	0.72	62
weighted avg	0.83	0.82	0.81	62

Figure 5.24 Classification metrics of SVM (upper face)

	precision	recall	f1-score	support
angry	0.75	1.00	0.86	9
disgust	1.00	1.00	1.00	12
fear	0.67	0.40	0.50	5
happy	0.93	1.00	0.96	13
sad	1.00	0.50	0.67	6
surprise	0.89	0.94	0.91	17
accuracy			0.89	62
macro avg	0.87	0.81	0.82	62
weighted avg	0.89	0.89	0.88	62

Figure 5.25 Classification metrics of SVM (full face)

Figure 5.26 Classification metrics of upper (above) and full face (below)

unconfident samples that are passed from stage-1 are viewed by the multi-feature fusion model. The accuracies obtained at different stages in Sternnet are shown in the table 5.7. The overall normalized confusion matrix and the classification metrics of overall SternNet is shown in the tables 5.9 and 5.8. The designed SternNet can also be used in other applications and the identification of RoC helps in complicated object classification tasks. This is a very important aspect to be considered since the present neural network models do not measure the confidence and credibility of their prediction.

5.5 Comparision of the proposed model with other existing FER models

The proposed model can be compared with other existing FER models which have a similar method of sample extraction and testing procedure. The proposed model as seen in the experiment outperforms the VGG16 model in terms of classification accuracy and provides better results in other classification metrics. The proposed model uses the dataset that has only one frame per labeled sequence and also is subject-independent. There exist many models which have included more than one sample (last three or five

Table 5.9 Confusion Matrix of SternNet model after five-fold-cross-validation

Emotions	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	**0.978**	0	0	0	0.022	0
Disgust	0	**1**	0	0	0	0
Fear	0	0	**0.92**	0.04	0	0.04
Happy	0	0	0	**1**	0	0
Sad	0	0	0	0	**1**	0
Surprise	0	0	0	0	0	**0.988**

Table 5.10 comparision of the proposed model with other existing standard FER models

S.No	Methodology	Number of Classes	Accuracy	Validation Method
1	Simultaneous facial feature [127] tracking	6	94.04%	Leaveone-out cross validation
2	An enhanced independent component-based [128]	6	93.23%	Leave one-out cross validation
3	Graph-preserving sparse nonnegative matrix factorization [129]	6	94.30%	Leave one-out cross validation
4.	Manifold structure learning using coordinates [130]	6	94.31%	Leave one-out cross validation
5	Spatial-Temporal Motion LBP and Gabor Multi-orientation Fusion Histogram [131]	6	95.8%	Leave one-out cross validation
6	Fusing Gabor and Local Binary Pattern Features [132]	6	96.5%	10-fold cross-validation scheme
7	Local Binary Patterns [75]	6	84.5% Tree-Augmented-Naive Bayes (TAN) classifiers	5-fold cross-validation scheme
		6	91.5% SVM (linear) LBP	10-fold cross-validation scheme
		6	92.6% SVM (RBF) LBP	10-fold cross-validation scheme
8	Hierarchical Deep Neural Network Structure [133]	6	96.46% (last three frames from sequences were used)	10-fold cross-validation scheme
9	Using Features of Salient Facial Patches [134]	6	94.1%	10-fold cross-validation scheme
10	The proposed work (SternNet)	6	98.1%	5-fold cross-validation scheme

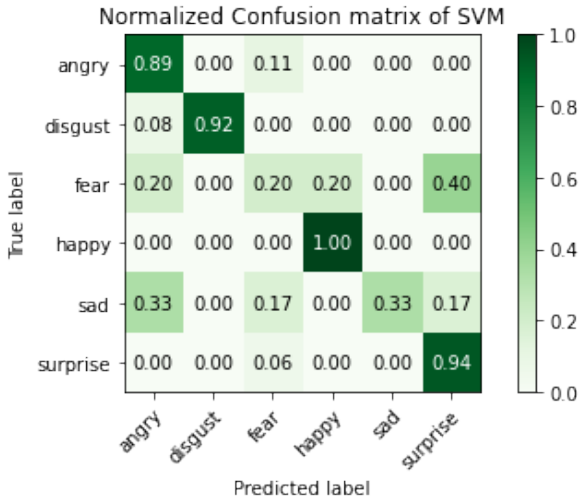

Figure 5.27 Normalized confusion matrix of SVM (upper face)

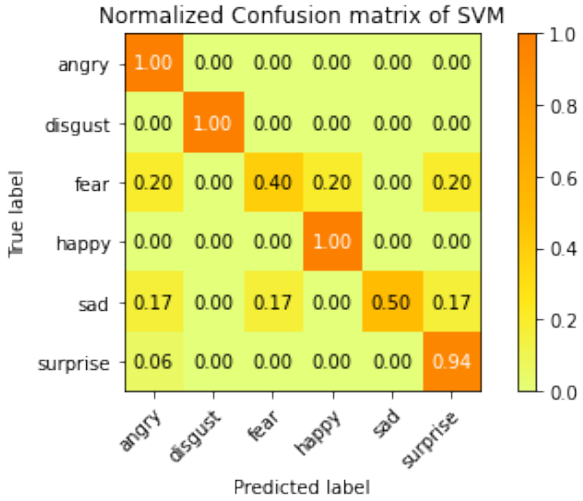

Figure 5.28 Normalized confusion matrix metrics of SVM (full face)

frames) per labeled sequence. The drawback of considering more than one frame per sequence is the bias as a model can similar frames in testing and training datasets. The important parameters that were selected in the comparison are methodology, extraction of training and testing samples from the dataset, number of classes used in the model, number of frames, and testing method. The table 5.10 compares SternNet with other existing FER models and the results conclude that the proposed model achieved better classification accuracy when compared to other models.

5.6 Conclusion

The proposed model SternNet is multi-stage facial expression recognition model that is built on the basics of the rank of confidence. The rank of confidence is a new metrics proposed in the research work analyses the face in different spatial regions and measures the confidence of the prediction. The RoC of rank 1 has a high confidence score when compared to other samples having RoC-2 and RoC-3. The confident samples are classified in stage-1 of SterNet and other samples are moved to stage-2 for classification. The concept of RoC can be applied not only to FER applications but also in other classification problems. The proposed model has achieved better performance when compared to other existing FER models in terms of classification metrics.

Chapter 6

VGG-Face and LSTM based Deep Neural Network for Near Infrared Facial Expression Recognition

6.1 Introduction

Facial expression recognition (FER) seeks to distinguish and categorise the meaningful emotions of numerous facial muscles into discrete emotion categories. An important premise in FER is that humans universally show six primary expressions: pleasure, sorrow, surprise, fear, anger, and disgust, as influenced by Ekman [135]. FER has been the subject of several investigations due to its potential applications in the realm of human–computer interaction [136]. Furthermore, facial expression recognition has a wide range of possible applications, including the service business, criminal investigation and interrogation, medical assistance, and so on.

There exist numerous works that can detect facial expressions in visible light. FER models built under Uncontrolled visible (VIS) light (380–750 nm) have common issues since in ambient circumstances VIS images can vary with place and time, resulting in considerable differences in visual appearance and texture. The facial expression identification algorithms established thus far operate well under controlled conditions, however variations in lighting or light angle pose issues for the recognition systems are major issues in FER[1]. To satisfy the needs of real-world applications, facial expression detection should be achievable under varied lighting situations, including near darkness. Unfortunately, algorithms are sophisticated and not particularly dependable, for example, employing

the same preprocessing for different lighting orientations may not produce satisfactory results, and under favourable lighting circumstances, such preprocessing might lose vital information. So, there is a need to search for other solutions. The Near Infrared range (NIR) spectrum's longer wavelengths can penetrate haze, light, fog, smoke, and other atmospheric conditions better than visible light. This frequently results in a crisper, less distorted image with higher contrast than what can be seen with visible light for long-distance imaging. The current deep FER standard is RBG or grey data, although these data are sensitive to lighting conditions. In contrast, infrared images that capture the temporal dispersion of the skin produced emotions are not affected by changes in lighting, which is a promising substitute for studying facial expressions

In the proposed work the authors have designed a FER model that can classify facial expressions in Near Infrared videos and images. NIR is very similar to human perception but eliminates the color wavelengths, resulting in most objects seeming very similar to a black and white image. The variations in reflectivity of some objects, along with reduced atmospheric haze and distortions in the NIR wavelength enhance the facial details and vision at long ranges. Other imaging systems like thermal energy, show faces quite differently from visual perception. NIR images appear similarly to visible light, which means it can see things like faces and other written information on signs like visible light.

The majority of facial expression recognition systems use two parts in the process of recognizing emotions. The two steps are feature extraction and classification. Feature extraction is the process of analyzing facial images and extracting the image's possible characteristics. Because various expressions have varied facial expression properties, useful prospective features for enhanced classification may be obtained. Effective expression recognition, on the other hand, remains challenging. Various hand-crafted characteristics (e.g., HoG [137], LBP [138], SIFT [139]) are often used in conventional FER. However, most of these qualities are incapable of simultaneously examining all aspects. The convolutional neural network (CNN) has recently been used to FER with excellent results [140]. This is due to its high representation capabilities. The proposed work in this research work uses sequence-based design for building a FER model. There are prior studies [53], [51] focused on sequence-based techniques, in which an expression is expressed as a sequence

of frames with known time stamps, to consider temporal information. The combination of convolutional and recurrent neural networks (CNN–RNN) (or CNN-LSTM) frameworks has emerged as a preferable solution because recurrent neural networks (RNN) perform well in processing diverse sequences with contexts. RNN usually uses the characteristics retrieved by CNN as input and later encodes temporal aspects. These CNN–RNN frameworks may leverage the benefits of CNN and RNN to concurrently learn appearance characteristics and temporal information, producing greater predictive performance than typical image-based FER approaches.

6.2 Related Works on Facial Expression Recognition models

FER has been investigated in the computer vision field for decades [47], [48]. According to the existing FER models, the approaches may be divided into two categories that are Static image-based and Dynamic image sequence-based approaches. In recent years, the usage of deep learning approaches in different computer vision challenges has risen. In pattern recognition tasks, Deep Neural network (DNN) models such as convolutional neural networks and recurrent neural networks are often utilised. In this part, we will go through some of the existing FER models.

Zhang et al. [49] suggested one of the still image-based approaches on the FER-2013 Challenge [50], the authors utilised a deep CNN accompanied by a linear one-vs-all support vector machine (SVM) and obtained good classification accuracy. Yu et al. [51] suggested an emotion detection module based on an ensemble of several networks, each with a separate set of weights. Breuer and Kimmel [52] investigated the potential of DNNs to grasp emotions by evaluating several CNN visualisation methodologies. Jung et al. [53] improved FER accuracy by employing two separate CNN models. Zhao et al. [10] defined the deep region using multi-label learning (DRML), which uses feed-forward networks to understand facial regions and assess structural patterns of the face by forcing knowledge to be captured by learnt weights. Mollahosseini et al. [54] suggested a network with two convolutional layers, each with max-pooling, and four Inception layers between them. The network is a one-component architecture that takes in captured facial images and categorizes them into six basic expressions along with neutral expressions. The FER

approach employs neural networks that leverage pre-trained networks, and models are deployed to save training time. The goal of using these pre-trained networks is to employ weights that have been developed during training on huge datasets such as Imagenet [33]. Kahou et al. [55] revealed the advantages of using pre-trained networks. It is worth noting that the temporal relationship between image frames in the sequence is critical for detecting face emotions. Recently, there has been a greater emphasis on methods that capture spatial–temporal aspects [[56], [57], [58]]. For video-based expression recognition, Liu et al. [59] employed a 3D-CNN architecture. They suggested a CNN architecture with flexible facial action parts model constraints that can learn spatial–temporal properties as well as locate facial action parts. For FER, Khorrami et al. [60] built a CNN–RNN architecture. They also looked at how much each network adds to the framework. Jaiswal et al. [58] proposed a model for obtaining temporal information using a mixture of CNN and BiLSTM, which outperformed other models in terms of accuracy. Fan et al. [57] developed a hybrid network that extracted features using a 3DCNN architecture and then utilised RNN to capture the temporal relationships for FER. According to the preceding discussion, multiple network integration and CNN–RNN frameworks considerably increase FER performance. The objective of the proposed work is to learn discriminative spatial–temporal features, particularly temporal motion context information using VGG-Face CNN-LSTM based architecture.

6.3 Deep Neural Networks

This section describes the proposed FER model using VGG-Face and LSTM neural networks. The following topics discuss the implementation and preprocessing steps involved in designing a spatio-temporal-based FER model.

6.3.1 Implentation of VGG-Face CNN model using transfer learning

Deep CNNs are robust models capable of capturing useful spatial information. Overfitting is a difficult problem in machine learning that happens while utilising smaller datasets. The problem of overfitting is magnified in deep neural networks since they include millions of parameters through several hidden layers. Using a fine-tuned CNN

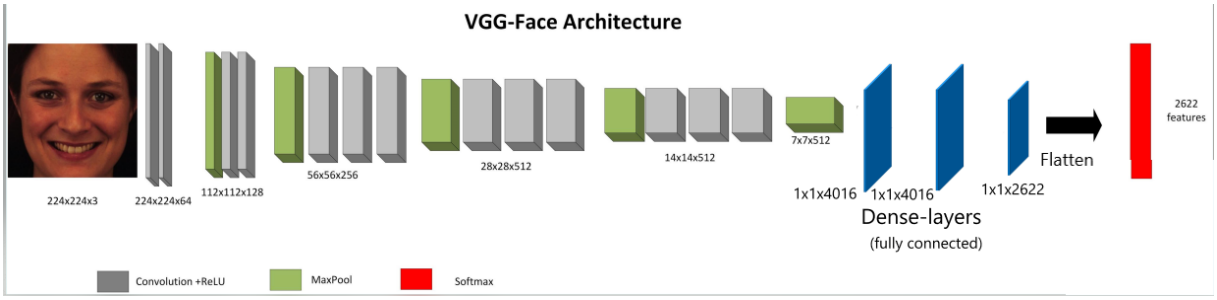

Figure 6.1 Architecture of VGG-Face

model instead of training the CNN model on smaller datasets improves CNN generalization. Fine-tuning is a transfer learning strategy that focuses on maintaining and transferring knowledge gained while solving one problem to a different but related challenge. Because CNNs are made up of numerous layers and a large number of parameters, using pre-trained weights derived from larger databases through fine-tuning could help prevent overfitting. When trained on smaller datasets, CNN models are more likely to catch patterns that are specific to the dataset. Because of this behavior, the model may become over-fit and difficult to generalize to additional external data. In this work, we employ the VGG-face mode [141], which yielded state-of-the-art results on the LFW [89] and YFT [142] databases. As illustrated in Figure 6.1, each convolutional layer is followed by a rectification layer, and each convolutional block ends with a max pool layer. Since the VGG- Face was built for face recognition, it can collect crucial low and high-level characteristics on the face. Table 6.1 compares several CNN models and their applications, and the VGG-Face model was chosen for this work since it has validated its efficiency on well-established datasets such as Labeled Faces in the Wild and YouTube Faces.

The RGB image may be directly fed into the CNN. The model applies filters to the image, resulting in feature maps. The kernel (filter) size, number of filters, and kind of padding are the critical hyper-parameters to consider in CNN. The hyper-parameters determine the depth of the output feature map. Consider **I**[j,k] as the input picture, bf H as the kernel, and **O** as the output matrix after convolution. The equation **??** gives a way for determining the size of the image after convolution. CNN employs several padding methods. CNN's numerous padding techniques are valid, full, and the same. Valid-padding reduces the size of the output matrix after convolution. Whereas, the output matrix has the same size as the input matrix with the same padding. The stride

Table 6.1 Comparison of popular pretrained CNN models

Model	Authors	Year	Advantages	Applications and Datasets tested
AlexNet [86]	Krizhevsky et al.	2012	Application of Rectified Linear Units (ReLUs) as activation functions	Object Detection ImageNet [33]
VGG-16,19 [87]	Simonyan et al.	2014	Implementation of deeper convolutions through stacking uniform convolution layers	Object Detection ImageNet
VGG-Face [88]	Parkhi et al.	2015	Compution of VGG-Face CNN descriptors that model human faces	Face Recognition Evaluated on exclusive face related datasets like Labeled Faces in the Wild [89] and YouTube faces [142]
Inception V1 [91]	Szegedy et al.	2015	Implementation of convolutions with different kernel size filters through addition of inception modules	Object Detection ImageNet
Res-Net 50 [92]	Kaiming et al.	2015	Application of skip connections and Batch normalization	Object Detection ImageNet
MobileNet	Howard et al.	2017	Introducing streamlined architecture that uses depth-wise separable convolutions	Efficient model for mobile and embedded vision applications Evaluated on ImageNet, Fine Grained Recognition, Large Scale Geolocalizaton and Face attributes

approach is another important feature of convolution. The amount of the kernel's shift during the convolution processing step is referred to as striding. The kernel moves across the picture during the convolution process. The stride number specifies the shift's step size. We can raise the stride value to lessen the dimensions of the feature map or the repeated activity. Let n denote the image size, f the filter size, c the colour map depth (number of channels), p the employed padding, s the stride number, and z the number of filters. After convolution of the feature maps with the filters, Equation6.1 yields the resulting feature map dimensions.

$$\mathbf{O}[m,n] = (\mathbf{I} * \mathbf{H})[m,n] = \sum_j \sum_k \mathbf{H}[j,k] \mathrm{x} \mathbf{I}[m-j, n-k] \tag{6.1}$$

$$[n,n,c] * [f,f,c] = [[\frac{n+2p-f}{s} + 1], [\frac{n+2p-f}{s} + 1], z] \tag{6.2}$$

Following the convolution, the feature maps are connected to an activation function. The Rectified Linear Unit (ReLU) is the most often utilised activation function in computer vision issues. In order to provide non-linearity in the model, ReLU is also inserted

after the convolutional layer. Following the convolution, a special layer called the pooling layer is applied, which down-samples the features. The pooling layer decreases the size of the feature map by reducing the picture while keeping key features at the same time. The equation for determining the dimensions of feature maps is given in the equation 6.2

CNN pooling techniques that are often used include max-pooling and average-pooling. To obtain the few most efficient features from an input picture, the convolution-ReLU-max-pooling blocks are repeated. Finally, the feature maps are converted to 1-dimension using a flatten layer. VGG-Face, as seen in Figure 6.1, is a popular CNN model employed in this work.

6.3.2 Frame Aggregation

Precisely assessing per-frame input features does not produce adequate results since the frame of the subjects may vary in expression intensity. To boost performance, a frame aggregation scheme is employed. Many approaches for aggregating frames in each sequence have been proposed. These approaches can be divided into two categories: decision-level aggregation and feature-level aggregation. The most convenient method in decision-level aggregation is to simply concatenate the other frames to the input frame but, the number of frames in each sequence may vary. To construct a fixed-length feature vector for each sequence, two aggregation procedures have been investigated [143], [144]: frame averaging and frame expansion.

For feature-level frame aggregation in each sequence, statistical parameters like the average, max, average of squares, average of maximum suppression vectors, and so on are calculated for overall frames and are concatenated to the input frame.

In this proposed model we have used decision-level frame expansion since not all sequences taken from the Oulu CASIA dataset have a fixed number of frames. The Oulu CASIA NIR dataset has mainly sequences of frames under three lighting conditions (dark, weak, and normal). In this work, we have considered all the frames from various subjects under three lighting conditions. The consideration of spatio-temporal features can also help in designing a robust FER model that handles illumination variations. The number of frames in each sequence is investigated in the Oulu CASIA dataset. On calculating

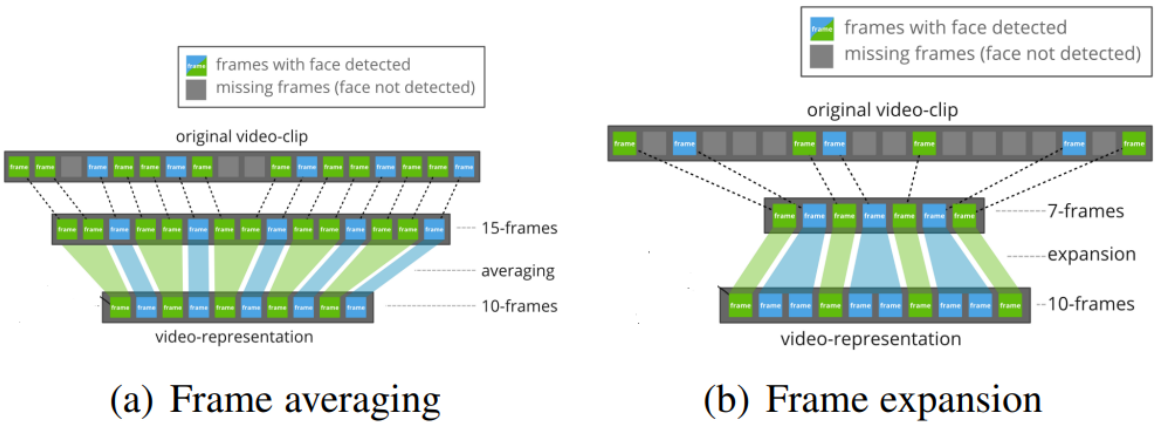

Figure 6.2 Frame Averaging and Frame Expansion of videos [2]

the number of frames in each sequence we have found that the least number of frames available in a sequence is 9. The highest number of frames available in a sequence is 47 and the mean number of frames in the Oulu CASIA sequences is 22. There can be an option of frame averaging by making fixed frames of 22 in all sequences but, we have considered a different method. When we remove frames in certain sequences which have frames greater than 22 there can be considerable temporal information loss. So to avoid this we have made 50 frames as a fixed number of frames in all sequences. Since the maximum frame length is 47, we need to expand the frame length in all the sequences. In the proposed work, we have stretched the frame length by adding only the last frame of the sequence till we achieve a frame length of 50. The last frame is added to the input in all the sequences since it expresses the highest intensity of emotion. The process of frame aggregation techniques is shown in the figure 6.2

6.3.3 Extraction of VGG-Face based deep convolutional feature vectors

The spatial features in each input frame are obtained using VGG-Face architecture, the input to the VGG-face consists batch of RGB images of faces displaying expressions in each sequence. The pixels of the input image are normalized and resized to the size of (224x224x3). As shown in Figure 6.1, the images are reshaped and given as input to the VGG-Face model. A sequence of (conv-relu-pool) layers is applied to the input images, and the resulting feature maps are flattened to generate 2622 parameters. The 2622 features

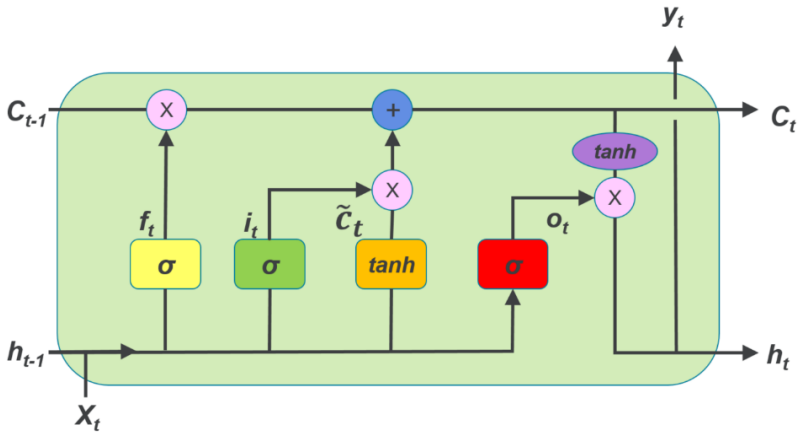

Figure 6.3 Architecture of LSTM [4]

are self-learnt features from the VGG-Face. The weights of the VGG-Face are imported through transfer learning. The weights of the VGG-Face model are imported from the data provided by the Visual Geometry Group, University of Oxford [88]. The provided weights from the source are in MatLab format. MatConvNet is a MatLab toolbox for CNN, and the provided weights are MatLab compatible. In this experiment, the weights of VGG-Face are converted from MatLab format to Keras compatible using machine learning libraries. The detailed procedure of conversion is explained by Sefik [101]. The spatial features are extracted using VGG-face and now the temporal information in these features needs to be investigated.

6.3.4 Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) is a kind of Recurrent Neural Network that is particularly intended to keep the neural network output for a given input from declining as it passes through the feedback loops. These recurrent networks outperform other neural networks in pattern recognition, especially in time-series-related tasks. The memory of previous input is crucial for completing sequence learning problems, and long short-term memory networks outperform other RNN designs by addressing the vanishing gradient problem. The fundamental LSTM unit is made up of a storage unit and three control gates. The figure 6.3 depicts the structural unit of a LSTM cell.

The three control gates and a memory unit structure are described in detail below.

1. Input gate: LSTM's input consists of H_{t-1} and X_t, where H_{t-1} is the prior time's hidden state and X_t is the recently received data at the present time..

$$I_t = \sigma\left(X_t W_{xi} + H_{t-1} W_{hi} + b_i\right) \tag{6.3}$$

2. Forget gate: The forgetting gate regulates the forgetting of the present node's past information. It may choose the network-remembered information and trigger the function through a sigmoid. This allows for the removal of superfluous and duplicate historical data. If the function's result is near 1, it signifies that the memory's information value is greater than prior data, and it will try to retain the present information value for the following step. If the function value is near zero, it is demonstrated that the memory information value is less than the prior value, and it will discard the majority of the memory information.

$$F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right) \tag{6.4}$$

3. Output gate: The gate primarily regulates the data output of nodes. If the node information reflects the critical feature, the output effect will be boosted, if it does not represent the critical feature, the output information will be lowered. Meanwhile, it may identify the output of the previous memory update to influence the magnitude of the next information output.

$$O_t = \sigma\left(X_t W_{xo} + H_{t-1} W_{ho} + b_o\right) \tag{6.5}$$

4. Memory cells: The goal of these memory cells is to save state information, i.e. to keep long-term historical data. Candidate memory cells must be estimated first in the initial stage. The computation procedure is identical to that of the previous three gates. The distinction is that the tanh function which has a range in [-1,1] is employed as an activation function. The below formula is the procedure for calculating candidate memory cells at time step t. The flow of information in the hidden state is then regulated by the input gate, forget gate, and output gate. The

computation of the present time step memory cell incorporates upper time step information and the present state memory cell. The calculation of current time step memory involves both memory cell and current time step candidate memory cell which are regulated through the forgetting gate and input gate present.

$$\tilde{C}_t = tanh\left(X_t W_{xc} + H_{t-1} W_{hc} + b_c\right) \tag{6.6}$$

$$C_t = F_t \otimes C_{t-1} + I_t \otimes \tilde{C}_t \tag{6.7}$$

6.4 Design of Spatio-Temporal Deep Convolutional RNN FER model

The proposed model is designed using VGG-Face and LSTM which are discussed in previous sections. The initial module of the model does frame length expansion to tackle the frame imbalance issue. After the frame aggregation process, the length of all sequences is changed to 50. The obtained frames are pre-processed which includes operations like face detection, reshaping, pixel normalization, and resizing. In the next module of the proposed model, the spatial features of all the frames (collection of 50 frames) in the given sequence are derived through the VGG-face CNN. The filter weights are loaded from pre-trained model weights through transfer learning. The weights of the VGG-Face are frozen and not trained because it already possesses rich discriminative features for detecting facial traits as it has been extensively trained on face-related datasets. The feature maps obtained from VGG-Face are shown in the figures 6.9 The application of VGG-Face on each frame results in forming 2622 deep learnt features. So from each frame, we obtain 2622 deep features and for each sequence, we obtain 13100 (2622*50) deep features. The important facial features are now obtained using VGG-Face and now it is necessary to investigate the temporal information in these consecutive deep learned feature vectors. We have implemented a 3000 LSTM cell sequence in the next layer. The following layers after the LSTM layer include dense and dropout layers. The final 256 features from the LSTM are connected with 256 spatial features taken from the last time

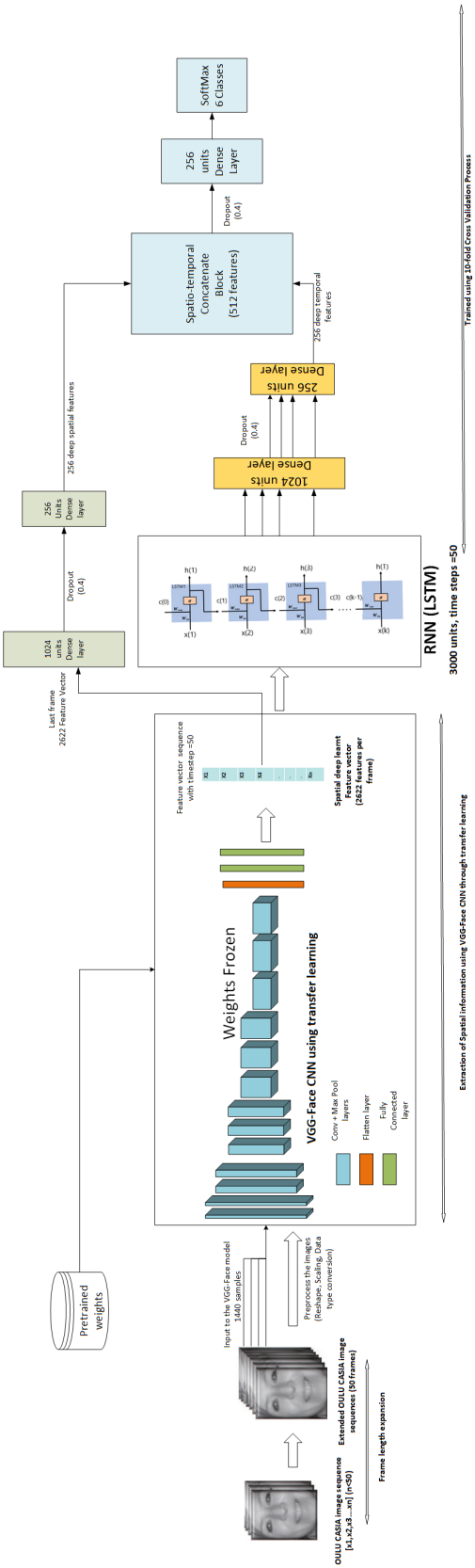

Figure 6.4 Architecture of proposed Spatio-Temporal FER model using VGG-Face and LSTM

Figure 6.5 Anger, disgust, fear, happiness, sadness, surprise images of one person from Oulu CASIA NIR facial expression database [5], [6]

frame (50th time frame). The addition of spatial features of the recent time frame to the temporal feature vector helps the model in recognizing recent facial expression variations which help in recognizing emotions. The final part of the architecture is connected with a softmax layer of 6 outputs. The designed FER model is capable of classifying facial expressions into six emotions. Angry, disgust, fear, happy, sad, and surprise are the six emotions that the designed FER model in this work can identify.

6.5 Datasets

Oulu-CASIA NIR and VIS facial expression database: The Oulu-CASIA NIR and VIS facial expression database has 80 participants aged 23 to 58 years old, with six expressions (surprise, happiness, sadness, anger, fear, and disgust). The six expressions of one subject are shown in the figure 6.5. The dataset contains 73.8 percent of the total subjects as male. The entire database is divided into two sections, one of which was shot in Oulu by the Machine Vision Group of the Oulu University which has 50 students, the majority of them are Finnish people. The other was taken in Beijing by the National Chinese Academy of Sciences, Pattern Recognition Laboratory consisting of 30 persons, all of whom are Chinese. The subjects were requested to sit on a chair in the observation room in such a way that he/she could see the camera. The camera-to-face distance is around 60 cm. All expressions are taken under three distinct lighting conditions. Normal, weak, and dark environments Normal lighting implies that there is enough illumination. Weak lighting indicates that just the computer display is turned on, and the individual is seated in front of the computer on a chair. Dark illumination denotes near total darkness. The figure 6.6 shows how effectively face characteristics like wrinkles and furrows may be observed in NIR and VIS pictures. The images show in the figure represent the same

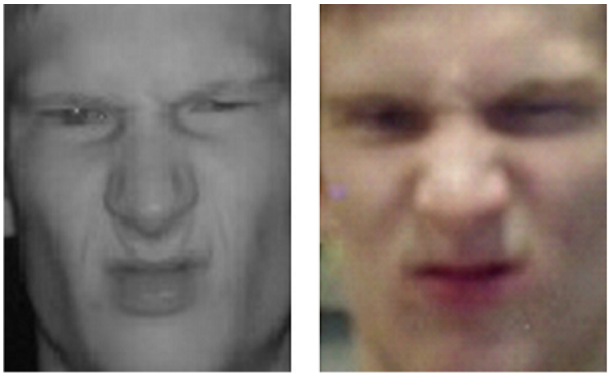

Figure 6.6 Comparison of NIR (left) and VIS image from Oulu CASIA [5]

frame from the image sequence. The contours of the facial features are clearly outlined in NIR image when compared with VIS image. There are no dark corners in the NIR image. However, certain dark patches induced by self-occlusion may be seen in the VIS image.

In the proposed work we have considered all the images from three different illumination variations of NIR images. The total number of sequences present in each illumination variation is 80 so, we have obtained a total of 240 sequences of 80 subjects taken under three distinct lighting conditions. The major issue in the dataset is the frame length imbalance observed in all the sequences. On calculating the number of frames in each sequence we have found that the least number of frames available in a sequence is 9. The highest number of frames available in a sequence is 47 and the mean number of frames is 22. This imbalance issue is solved using the frame expansion technique as discussed in the previous section. After the process of frame aggregation, the number of frames in all sequences is 50. The total number of images considered in the proposed work is 12000 (Number of sequences in each illumination variant x Total number of illumination variants x Number of frames present in all the sequences = 80x3x50). We have considered subject independent classification which means the entire frames belonging to a sequence of a subject will be either in the training fold or testing fold. We have implemented a ten-fold cross-validation scheme to calculate the accuracy of the model.

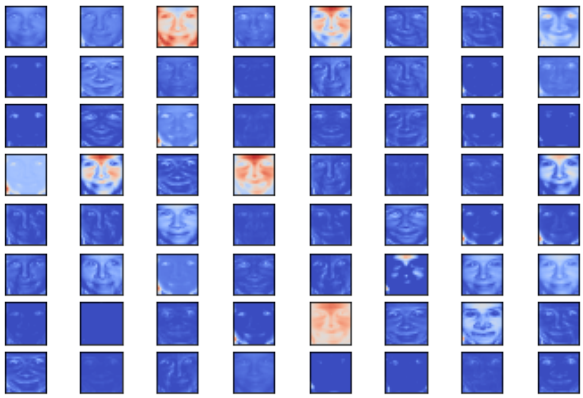

Figure 6.7 Lower-level feature-maps of CNN extracting minor details of facial characteristics like lines, curves, and dots

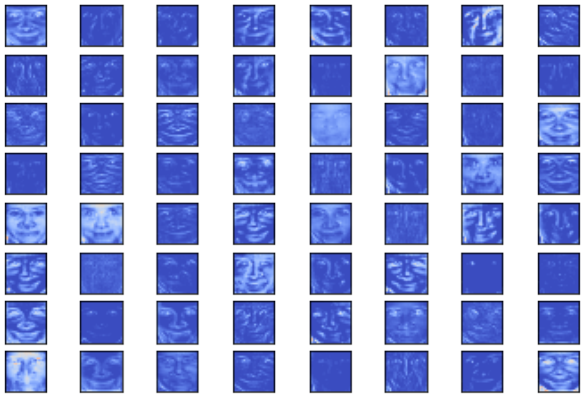

Figure 6.8 Lower-mid level feature-maps of CNN extracting basic facial details at important landmark points

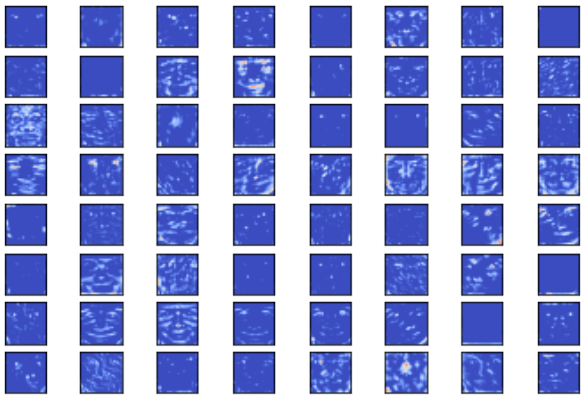

Figure 6.9 Higher-mid level CNN feature-maps detecting facial texture and patterns at eyes, nose, and mouth

6.6 Results and Discussion

6.6.1 Training the proposed FER model with facial expression datasets

We trained up the FER model with Google's Colaboratory (colab). Google Colab provides Nvidia 1xTesla K80 GPU with 2496 CUDA cores. Colab provides 12GB of GDDR5 VRAM. Colab also provides a single-core hyper-threaded Xeon Processor with a clock speed of 2.3GHz.The procedures for preprocessing the input frames are discussed as in the prior segment. Keras deep learning framework is used to create the deep neural network architecture. In this model, an error function called categorical cross-entropy is applied. The formula for calculating categorical loss is given in equation 6.9. In this experiment, we utilised the ADAM [103] optimizer since it integrates the impacts of the RMSProp [104] (the capacity to cope with non-stationary objectives) and ADAGrad [105] optimizers (the ability to deal with sparse gradients). Individual adaptive learning rates for various parameters are determined in ADAM using estimations of the gradient's first and second moments. ADAM has several benefits, including the fact that the magnitudes of parameter updates are invariant to gradient rescaling, that the stepsizes are controlled by the step-size hyperparameter, and that it does not need a stationary objective. The authors also explored the efficacy of ADAM in multi-layer neural networks and deep CNNs. The studies presented in the work demonstrate that ADAM is stable and well-suited to a wide range of non-convex optimization problems in machine learning. Table 6.2 shows the equations and other parameters utilised by the ADAM optimizer. The softmax layer at the bottom of the proposed model generates the predicted emotion values for all input images. The softmax layer is a function that converts randomly generated data into a suitably ordered probability distribution. The Softmax layer function's output differs from (0,1). There are six classes in the proposed FER model. Let us consider t_i and y_i be the target and the softmax score of the i^{th} class of a sample. The softmax activation function is explained by equation 6.8, for each class i, there exists a softmax score according to equation 6.8. The class with the highest softmax score is predicted as the output class of the respective sample.

$$\textit{Softmax score for each class i= 1 to 6} : f(y)_i = \frac{e^{y_i}}{\sum_j^{N=6} e^{y_j}} \tag{6.8}$$

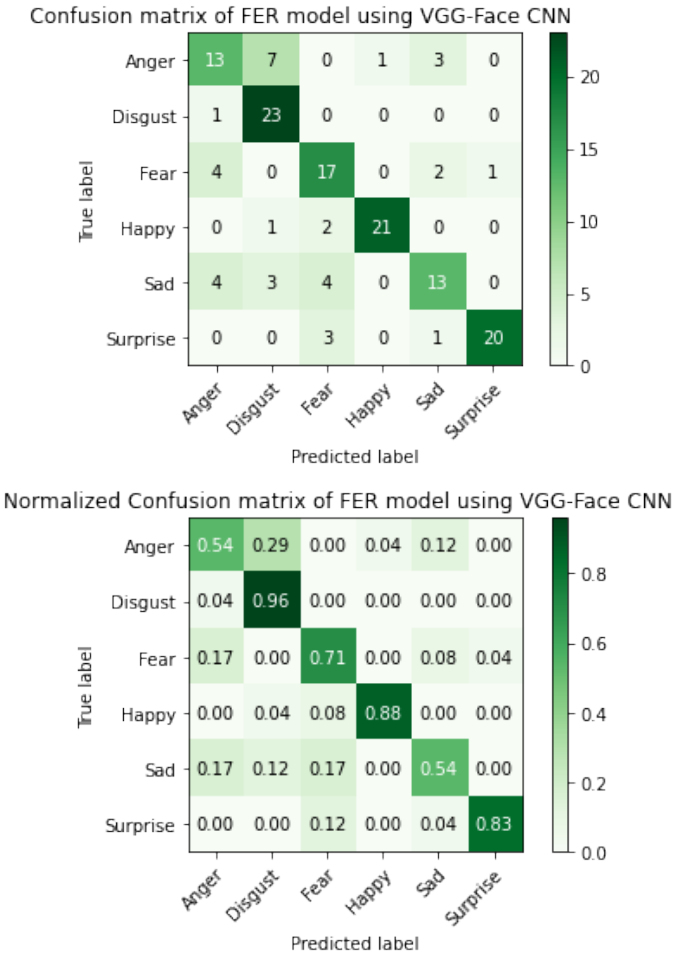

Figure 6.10 The confusion matrix plots of fine-tuned VGG-Face model on Oulu-CASIA dataset

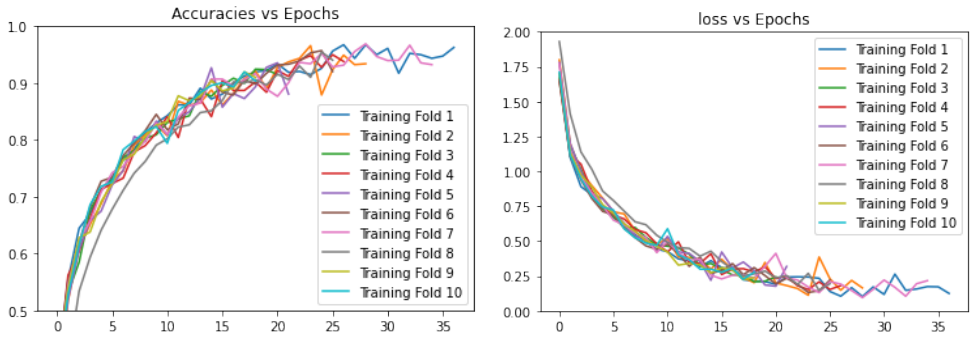

Figure 6.11 The Accuracy vs Epochs (left) and Categorical cross entropy loss vs Epochs (right) plots during the training (10-fold cross-validation) of Oulu-CASIA dataset on the fine tuned VGG-Face FER model

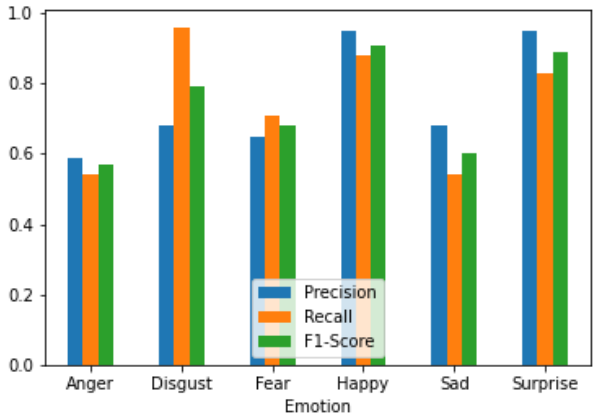

Figure 6.12 Comparison of Precision, Recall, and F1- score information of the fine-tuned VGG-Face model on NIR Oulu CASIA dataset

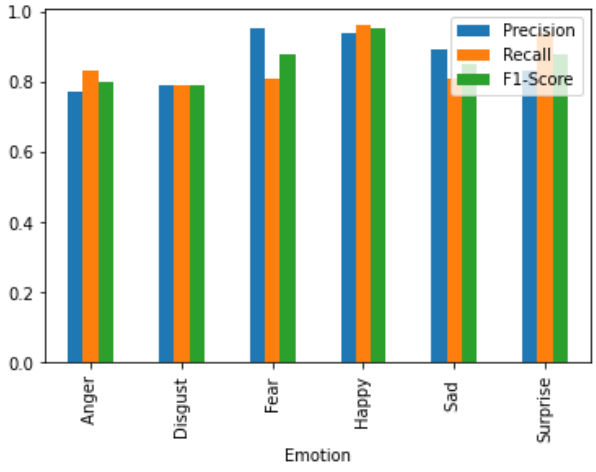

Figure 6.13 Comparison of Precision, Recall, and F1- score information of the proposed spatio temporal model on NIR Oulu CASIA dataset

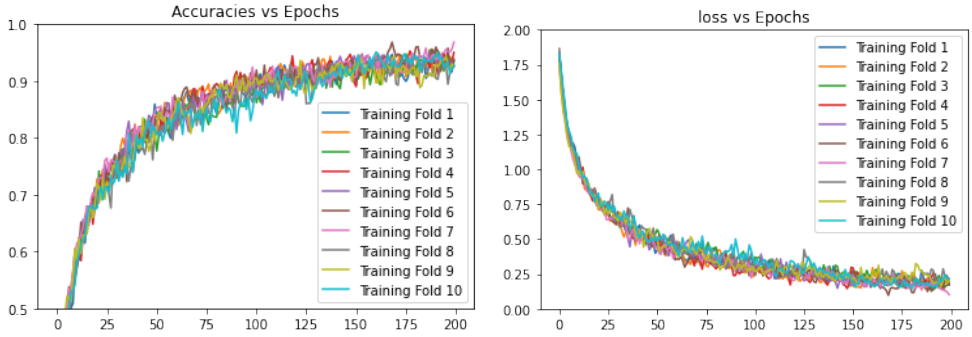

Figure 6.14 The Accuracy vs Epochs (left) and Categorical cross-entropy loss vs Epochs (right) plots during the training (10-fold cross-validation) of Oulu-CASIA dataset on the proposed Spatio-temporal VGG-face RNN based FER model

$$Categorical\ Cross\ entropy\ error: -\sum_{i=1}^{N=6} t_i log(y_i) \tag{6.9}$$

The below equations explain the procedure to update weights using the ADAM optimizer. Table 6.2 explains the values of the parameters used in ADAM :

$$\theta_{t+1} = \theta_t - \frac{\eta \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t},\ \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

and where

$$m_t = (1-\beta_1)g_t + \beta_1 m_{t-1}$$

$$v_t = (1-\beta_2)g_t^2 + \beta_2 v_{t-1}$$

$$g(gradient) = \nabla J(\theta_{t,i})$$

The sum of all outputs from the softmax layer equals one. In Multi-Class classification problems, the targets are one-hot encoded, making only the positive class appear in the categorical loss function.

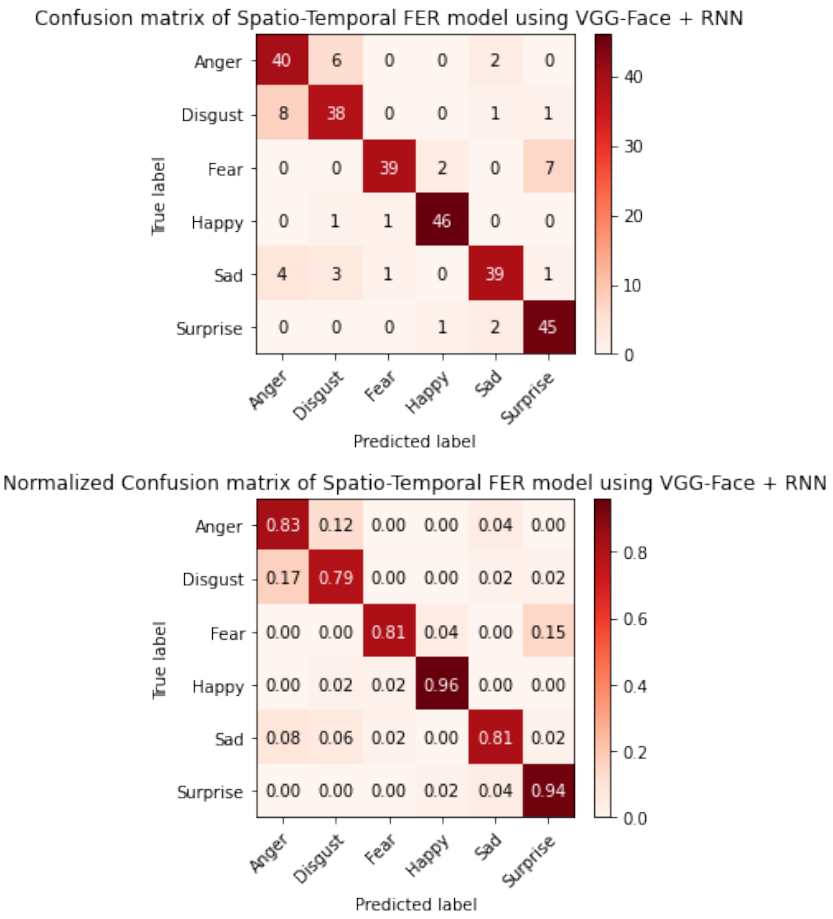

Figure 6.15 The normalized confusion matrix plots of the proposed Spatio-temporal FER model on Oulu-CASIA test data

Parameter	**Value chosen**	**Role in ADAM**
Epsilon ,ϵ	10^{-8}	preventing division by zero
Learning rate,η	0.001	step size in each iteration
First momentum,β_1	0.9	speed of convergence
Second momentum,β_2	0.99	speed of convergence

Table 6.2 Values of different parameters used in the ADAM optimizer

Table 6.3 Comparison of accuracy rates of three models on Oulu-CASIA dataset

Model	**Dataset**	**Accuracy(10-fold cross-validation)**	**Method**	**Number of Frames in each Sequence**
Fine-tuned VGG-Face CNN	Oulu-CASIA	74.3%	10-fold-cross-validation	1 (last frame)
Spatio-temporal VGG-Face and LSTM model	Oulu-CASIA	78.3%	10-fold-cross-validation	5 frames (Frame Skipping)
Spatio-temporal VGG-Face and LSTM model	Oulu-CASIA	85.6%	10-fold-cross-validation	50 frames (Frame aggregation)

6.6.2 Classification Metrics

The model's essential evaluation metrics discussed in this work are accuracy, precision, recall, and F1 score. Let TP represents True Positives, FP represents False Positives, FN represents False Negatives, and FP represents False Positives.

1. **Accuracy**: Accuracy (Acc) is useful for evaluating model efficiency. However, when there is a class imbalance problem, it is essential to consider other critical metrics, such as precision and recall.

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{6.10}$$

2. **Precision**: Precision (P) underlines the ability of the model to select the class of choice. P is based on TP and FP. False Positives are the number of predictions that the model misclassifies as positive when the true label is negative.

$$P = \frac{TP}{TP + FP} \tag{6.11}$$

Table 6.4 Comparison of classification accuracy rates of the proposed Spatio-Temporal Feature-based VGG-Face LSTM model using 10-fold-cross-validation scheme

Author	Methodology	Number of classes	Illumination Type	Accuracy	Validation Method
Zhao et al. [145]	LBP-TOP descriptors and support vector machine	6	Average of three (Normal, Weak, and Dark)	69.5%	10-fold-Cross-Validation
Zhao et al. [145]	LBP-TOP descriptors and sparse representation classifier	6	Average of three (Normal, Weak, and Dark)	74.11%	10-fold-Cross-Validation
Wu et al. [6]	Three-Stream 3D Convolutional Neural Network	6	Dark illumination condition.	78.42%	10-fold-Cross-Validation
Chen et al. [146]	Three-Stream Convolutional Neural Network with Squeeze and Excitation block SETFNET and SETFNET global	6	Dark illumination condition.	80.34% 81.67%	10-fold-Cross-Validation
Jung et al. [147]	Deep Temporal Geometry Network Deep Temporal Appearance network	6	Normal illumination	74.17% 74.38%	10-fold-Cross-Validation
Jung et al. [147]	Deep temporal Appearance-geometry network (DTAGN) Weighted	6	Normal illumination	80.62%	10-fold-Cross-Validation
Jung et al. [147]	DTAGN(Joint)	6	Normal illumination	81.46%	10-fold-Cross-Validation
Jeni et al. [148]	precise 3D shape registration + Constrained Local Models (CLM) + Multi-class SVM	6	Average of three (Normal, Weak, and Dark)	69.25%	10-fold-Cross-Validation
Our Proposed model	VGG-Face based Spatio-temporal Deep Neural Network	6	Average of three (Normal, Weak, and Dark)	85.6%	10-fold-Cross-Validation

3. **Recall**: Recall (R) is the other classification metric that conveys the ability of the model to predict all classes of interest in the dataset. R is based on TP and FN. FN is the number of predictions that the model misclassifies as negative when the true label is positive.

$$R = \frac{TP}{TP + FN} \tag{6.12}$$

4. **F1 Score**: Good precision and recall must be preserved for every model. A good classifier aims to choose the correct class without any error (precision) and, at the same time, to choose as many correct classes as possible (recall). A successful trade-off between precision and recall must be preserved. The F1 score offers a good combination of two measures of recall and precision. The F1 score is the harmonic mean of recall and precision.

$$F1\ Score = 2 * \frac{P * R}{P + R} \tag{6.13}$$

6.6.3 Comparison of the proposed model with other popular FER models

The proposed model is compared with other popular FER models that classify emotions from infrared videos. The comparison involves discussing the important parameters such as the illumination method considered, validation scheme, and chosen methodology. The table 6.4 explains the comparison of the proposed model with other prominent FER models. Zhao et al. [145] have used the handcrafted local binary pattern to extract features from the images. The method achieves an average accuracy of 69% and 74.11% using SVM and SRC classifiers respectively on three illumination methods. Jung et al. [147] have achieved an accuracy of 81.46% on the normal illumination method using a 10-fold-cross-validation scheme. Compared with other models, our proposed model outperforms in terms of accuracy by eliminating intra-class correlation by fusing spatial and temporal features.

6.7 Conclusion

The proposed FER model in this work extracts crucial patterns of facial expressions by fusing spatial and temporal features. The designed architecture takes frame aggregated sequences as input and utilizes transfer learning to import rich discriminative filter weights of the VGG-Face model. The obtained spatial features of each frame are concatenated and sent as input to an LSTM model. The LSTM model derives important temporal features in the sequence. The resultant 256 features are concatenated with the spatial features of the recent time frame. The final layer has a softmax layer of six classes which classifies six emotions. The comparison of classification metrics shows that the proposed model discriminates emotions better than a conventional CNN model. The results also conclude the proposed model outperforms other important FER models in terms of classification accuracy rate using a 10-fold-cross-validation scheme.

Chapter 7

Conclusions and Future Scope

This chapter concludes the thesis by underlining the main contributions. It also presents the possible directions of future work.

7.1 Conclusions

In the first contribution of the research work the classification of facial expressions using FACS and LogicMax has improved the accuracy rate of the FER by decreasing the influence of intra-class correlation of facial expressions. The performance of the model outperforms other state-of-the-art techniques in terms of various classification parameters on CK+ and Jaffe datasets. This work improves the precision of classifying emotions like Happiness, Disgust, and surprise by implementing a dual CNN architecture with logicMax. The dual CNN architectures are trained to classify facial expressions in upper and lower face regions. The LogicMax analyzes the predicted emotions found on the upper and lower face and decides the final class by selecting the most appropriate emotion. The proposed work can be extended by using other correlation methods on action units.

The second contribution in the thesis represents the use of multi-feature fusion design. The proposed FER model extracts crucial patterns of facial expressions using a combination of HOG and CNN features. Compared to a conventional convolutional neural network, the proposed FER model has a better discriminating ability in classifying similar emotions. In a traditional convolutional neural network, emotions like anger are strongly correlated with other emotions like neutral and sadness. The faces displaying emotions

like surprise are also often misclassified as fear by the CNN model. The proposed model succeeds in classifying similar emotions, which is a main drawback of the CNN model. The fusion of two different and complex features has shown faster categorical entropy loss convergence. The proposed hybrid model has reached good validation accuracy within a few epochs of training. The classification metrics like precision, recall, and f1-score convey that the proposed model has significantly improved discriminating expressions of sleepy and sadness on the Yale-Face dataset. The comparison of classification metrics and normalized confusion matrices show that the proposed FER model outperforms the other existing models in classifying facial expressions on the CK+, Yale-Face, and KDEF facial expression datasets. The proposed FER model has achieved accuracy scores of 98.11%, 97.84%, and 96.67% on CK+, KDEF, and Yale-face facial expression databases using the 10-fold-cross-validation process.

The third contribution in the thesis represents a novel SternNet model. The authors have designed a novel FER model known as SternNet that imparts stern rules in classifying facial emotions. The important aspect of this work is to improve classification efficiency by understanding the knowledge of rank of confidence. The rank of confidence measures how accurate is the prediction made by the FER model. The knowledge of rank of confidence makes the designers think about the credibility of the classification model. The proposed SternNet segregates the samples according to confidence and implements a two-stage methodology to predict subject-independent facial expressions. Combining human logical decisions in AI improves accuracy and also adds confidence to the prediction. SternNet can also be improved by adding more facial regions in stage-1 but it can also make the model more complex. The classification metrics noted in SternNet outperform existing FER models and also performed better than the conventional convolutional neural networks like VGG16.

The three contributions are modeled for static images and in the fourth contribution, we have used the dynamic frame model which takes an image sequence as input to the FER model. In the design, the model extracts crucial patterns of facial expressions by fusing spatial and temporal features. The designed architecture takes frame aggregated sequences as input and utilizes transfer learning to import rich discriminative filter weights of VGG-Face model. The obtained spatial features of each frame in the sequence are

concatenated and given as input to the LSTM model to comprehend the temporal features. The LSTM model derives important temporal features in the sequence. The resultant 256 features given by the LSTM branch are concatenated with the spatial features of the recent time frame (last frame of the sequence). The final layer has a softmax layer of six classes which classifies six emotions. The model is validated on Oulu CASIA Near Infrared Region sequence dataset on three illumination variations (normal, Weak and dark). The comparison of classification metrics shows that the proposed model discriminates emotions better than a conventional CNN model. The results also conclude the proposed model outperforms other important FER models in terms of classification accuracy rate using a 10-fold-cross-validation scheme and also shows good accuracy in low light scenarios.

7.2 Future Scope

The work proposed in this thesis can be extended for future research. The FER models also face issues in the pose angles of the faces. If the DNN models are trained on facial datasets that contain only frontal/posed faces, it is difficult to generalize faces on other multi-view facial expressions recognition. This problem can be solved using GANs where, the GAN-based FER models can predict even for multi-view FER. In future applications, we try to solve the ill-posed problem and address the serious view-based distortion encountered by the FER models. The next scope of the proposed FER models is to add the multi-head Attention modules. The concept of the transformer model is currently implemented on the image level through Vision Transformers which can further boost the accuracy rates of FER models. In the future, we can also add extra attention modules to improve the classification efficiency of the proposed models.

Publications

List of International Journals:

1. A.B. Ahadit and J. Ravi Kumar, "A Novel Dual CNN Architecture with LogicMax for Facial Expression Recognition," *Journal of Information Science & Engineering* , 37 (1): (2021). **(SCIE-Indexed, Published)**

2. A.B. Ahadit and J. Ravi Kumar, " A novel multi-feature fusion deep neural network using HOG and VGG-Face for facial expression classification," *Machine Vision and Applications*, 33(4), 1-23, (2022). **(SCIE-Indexed, Published)**

3. A.B. Ahadit and J. Ravi Kumar, "VGG-Face and LSTM based Deep Neural Network for Near Infrared Facial Expression Recognition," *Multimedia Tools and Applications*, (2022). **(SCIE-Indexed, Under Review)**

4. A.B. Ahadit and J. Ravi Kumar, "Real-Time Implementation of CNN based Drowsiness Detection System on NVIDIA Jetson Nano Platform," *SN Computer Science*, (2022). **(Scopus-Indexed, Under Review)**

List of International Conferences:

1. A.B. Ahadit and J. Ravi Kumar, "SternNet: A Rank of Confidence based Multi-Stage Facial Expression Classification," in *International Conference on Innovative Technologies in Recent Research (ICITRR), 2021* **(Published)**

Bibliography

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition*. Vol. 1, Ieee p: CVPR'05), 2005, pp. 886–893.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[4] D. Thakur, "Lstm and its equations," *https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af*, 2020.

[5] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[6] Z. Wu *et al.*, "Nirexpnet: Three-stream 3d convolutional neural network for near infrared facial expression recognition," *Applied Sciences*, vol. 7, p. 11, 2017.

[7] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[8] A. Mehrabian, "Communication without words," *Psychology today*, vol. 2, no. 4, 1968.

[9] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[10] 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.

[11] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from eeg using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2009.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *bmvc*, vol. 1, no. 3, 2015, p. 6.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[16] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[17] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[18] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 302–309.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[20] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[21] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48–57.

[22] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3359–3368.

[23] Y. Tang, X. M. Zhang, and H. Wang, "Geometric-convolutional feature fusion based on learning propagation for facial expression recognition," *IEEE Access*, vol. 6, pp. 42 532–42 540, 2018.

[24] Y. Wang, M. Li, C. Zhang, H. Chen, and Y. Lu, "Weighted-fusion feature of mb-lbpuh and hog for facial expression recognition," *Soft Computing*, vol. 24, no. 8, pp. 5859–5875, 2020.

[25] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. . Ren, "December). feature fusion of hog and wld for facial expression recognition," in *Proceedings of theIEEE/SICE International Symposium on System Integration*. IEEE, 2013, pp. 227–232.

[26] X. Xie and K. M. Lam, "Facial expression recognition based on shape and texture," *Pattern Recognition*, vol. 42, no. 5, pp. 1003–1011, 2009.

[27] D. T. Lin and D. C. Pan, "Integrating a mixed-feature model and multiclass support vector machine for facial expression recognition," *Integrated Computer-Aided Engineering*, vol. 16, no. 1, pp. 61–74, 2009.

[28] G. V. Reddy, C. D. Savarni, and S. Mukherjee, "Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features," *Cognitive Systems Research*, vol. 62, pp. 23–34, 2020.

[29] X. Pan, "Fusing hog and convolutional neural network spatial–temporal features for video-based facial expression recognition," *IET Image Processing*, vol. 14, no. 1, pp. 176–182, 2020.

[30] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint*, archivePrefix = arXiv, eprint = 1705.01842,.

[31] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in Proceedings of the IEEE international conference on computer vision *p*, 2015, pp. 2983–2991.

[32] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*," vol. 2016, pp. 3391–3399.

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

[34] S. E. Kahou, C. Pal, X. Bouthillier, and P. Froumenty, "G"ulçehre ç," in *Proceedings of the 15th ACM on International conference on multimodal interaction.* Vincent P, Courville A, Bengio Y, Ferrari RC, Mirza M, Combining modality specific deep neural networks for emotion recognition in video. p: Memisevic R, 2013, pp. 543–550.

[35] M. Koc, S. Ergin, M. B. G"ulmezoğlu, R. Edizkan, and A. Barkana, "Use of gradient and normal vectors for face recognition," *IET Image Processing*, vol. 14, no. 10, pp. 2121–2129, 2020.

[36] Liu, Ping and Han, Shizhong and Meng, Zibo and Tong, Yan, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[37] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via deep learning," in *IEEE international conference on smart computing.* p, 2014, pp. 303–308.

[38] 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.

[39] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW).* pages, 2015.

[40] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, pp. 4193–4203, 2017.

[41] A. R. Kurup, M. Ajith, and M. M. Ramón, "Semi-supervised facial expression recognition using reduced spatial features and deep belief networks," *Neurocomputing*, vol. 367, pp. 188–197, 2019.

[42] S. Datta, D. Sen, and R. Balasubramanian, "Integrating geometric and textural features for facial emotion classification using svm frameworks," in *Proceedings of International Conference on Computer Vision and Image Processing.* p, 2017, pp. 619–628.

[43] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *13th IEEE International Conference on Automatic Face and Gesture Recognition.* p, 2018, pp. 302–309.

[44] B. K. Kim, S. Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops *p*, 2016, pp. 48–57.

[45] M. S. Zia, M. Hussain, and A. Jaffar, M. Arfan, "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.

[46] S. F. Cotter, "Weighted voting of sparse representation classifiers for facial expression recognition," *IEEE 18th European Signal Processing Conference*, pp. 1164–1168, 2010.

[47] J. Chi, C. Tu, and C. Zhang, "Dynamic 3d facial expression modeling using laplacian smooth and multi-scale mesh matching," *The Visual Computer*, vol. 30, no. 6, pp. 649–659, 2014.

[48] M. Liu *et al.*, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[49] Z. Zhang *et al.*, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[50] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing. , , Heidelberg.* Berlin: Springer, 2013.

[51] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, vol. 34, no. 12, pp. 1691–1699, 2018.

[52] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," arXiv, preprint, 2017.

[53] H. Jung *et al.*, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[54] 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.

[55] S. E. Kahou, C. Pal, X. Bouthillier, and P. Froumenty, "G"ulçehre ç," in *Proceedings of the 15th ACM on International conference on multimodal interaction.* Vincent P, Courville A, Bengio Y, Ferrari RC, Mirza M, Combining modality specific deep

neural networks for emotion recognition in video. p: Memisevic R, 2013, pp. 543–550.

[56] A. Danelakis, T. Theoharis, and I. Pratikakis, "A spatio-temporal wavelet-based descriptor for dynamic 3d facial expression retrieval and recognition," *The visual computer*, vol. 32, no. 6, pp. 1001–1011, 2016.

[57] Y. Fan *et al.*, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016.

[58] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016.

[59] M. Liu *et al.*, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision. , Cham*, 2014.

[60] P. Khorrami *et al.*, "How deep neural networks can improve emotion recognition on video data," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016.

[61] 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.

[62] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7310–7311.

[63] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Scientific reports*, vol. 6, p. 22049, 2016.

[64] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[65] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

[66] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[67] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[68] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[69] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision.* Springer, 2016, pp. 425–442.

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

[71] Jung, Heechul, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim., "Joint fine-tuning in deep neural networks for facial expression recognition." in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2983–2991.

[72] A. S. Alphonse and D. Dharma, "Novel directional patterns and a generalized supervised dimension reduction system (gsdrs) for facial emotion recognition," *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 9455–9488, 2018.

[73] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

[74] M. Alif, A. Syafeeza, P. Marzuki, and A. N. Alisa, "Fused convolutional neural network for facial expression recognition," in *Proceedings of the Symposium on Electrical, Mechatronics and Applied Science*, 2018, pp. 73–74.

[75] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[76] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," *IETE technical review*, vol. 32, no. 5, pp. 347–355, 2015.

[77] 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.

[78] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The japanese female facial expression (jaffe) database," in *Proceedings of third international conference on automatic face and gesture recognition*, 1998, pp. 14–16.

[79] G. Corrado, "Scikit learn: Machine learning in python."

[80] P. Carcagnì and D. Coco, "Marco and leo, marco and distante, cosimo, facial expression recognition and histograms of oriented gradients: a comprehensive study," in Plus*, 4 (1), p*, 2015, pp. 1–25.

[81] B. Da and N. Sang, "Local binary pattern based face recognition by estimation of facial distinctive information distribution," *Optical Engineering*, vol. 48, p. 11, 2009.

[82] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, W. Gao, and A. Wld:, "robust local image descriptor," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1705–1720, 2009.

[83] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza, "Gender recognition from face images with local wld descriptor," in *2012 19th international conference on systems*. IEEE: signals and image processing (IWSSIP), 2012, pp. 417–420.

[84] F. Ahmed, E. Hossain, A. S. M. H. Bari, and A. S. M. Shihavuddin, "Compound local binary pattern (clbp) for robust facial expression recognition," in 2011 IEEE 12th

International Symposium on Computational Intelligence and Informatics (CINTI) *p*, 2011, pp. 391–395.

[85] J. Chen, Z. Chen, Z. Chi, and H. a. Fu, "and others, facial expression recognition based on facial components detection and hog features," *International workshops on electrical and computer engineering subfields*, pp. 884–888, 2014.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[87] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, type = preprint, archivePrefix = arXiv, eprint = 1409.1556, Tech. Rep., 2014.

[88] O. M. Parkhi, A. Vedaldi, and A. Zisserman, $vgg/software/vgg_face/$, University of Oxford, 2015.

[89] G. B. Huang, M. Mattar, T. Berg, and E. . Learned-Miller, *October)*. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Workshop on faces in'Real-Life'Images: detection, alignment, and recognition.

[90] L. Wolf, T. Hassner, and I. . Maoz, "June)," *Face recognition in unconstrained videos with matched background similarity*, pp. 529–534, 2011.

[91] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition *p*, 2015, pp. 1–9.

[92] 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 *p*, 2016, pp. 770–778.

[93] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos, "Cross-database facial expression recognition based on fine-tuned deep convolutional network," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. p, 2017, pp. 405–412.

[94] 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*. p, 2010, pp. 94–101.

[95] M. Pantic, M. Valstar, R. Rademaker, and L. . Maat, "July)," in *Web-based database for facial expression analysis*. InIEEE international conference on multimedia and Expo (pp. 5-pp). IEEE, 2005.

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

[97] D. Lundqvist and A. Flykt, "A. öhman, the karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, vol. 91, no. 630, pp. 2–2, 1998.

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

[99] A. Martinez and R. Benavente, "The ar face database, cvc," *Copyright of Informatica (0*, vol. 3505, p. 596, 1998.

[100] P. Viola and M. . Jones, "December). rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*. pp. I-I). Ieee: CVPR(Vol. 1, 2001.

[101] S. I. Serengil, "Matlab to keras conversion of weights," *https://sefiks.com/2019/07/15/how-to-convert-matlab-models-to-keras/*, 2020.

[102] P. Ekman, W. Friesen, and J. Hager, *Facial action coding system: Research nexus network research information*. UT: Salt Lake City, 2002.

[103] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization. arxiv," 2014, preprint.

[104] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop, coursera: Neural networks for machine learning," 2012, university of Toronto, Technical Report.

[105] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of machine learning research*, vol. 12, p. 7, 2011.

[106] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.

[107] J. C. Platt, N. Cristianini, and J. . Shawe-Taylor, "November)," *Large margin dags for multiclass classification*, vol. 12, pp. 547–553.

[108] C. Shan, S. Gong, and P. W. . McOwan, "September). robust facial expression recognition using local binary patterns," in *Conference on Image Processing(Vol*, I. International, Ed. pp. II-370). IEEE: 2, 2005.

[109] J. H. Friedman, "Another approach to polychotomous classification," 1996, technical Report, Statistics Department, Stanford University.

[110] G. P. Hegde and M. Seetha, "Subspace based expression recognition using combinational gabor based feature fusion," *International Journal of Image, Graphics and Signal Processing*, vol. 9, p. 1, 2017.

[111] S. Nigam, R. Singh, and A. K. Misra, "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain," *Multimedia tools and applications*, vol. 77, no. 21, pp. 28 725–28 747, 2018.

[112] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.

[113] L. Nwosu, H. Wang, J. Lu, I. Unwala, X. Yang, and T. Zhang, "Deep convolutional neural network for facial expression recognition using facial parts," *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big*

Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), IEEE, pp. 1318–1321, 2017.

[114] S. V. Y. R. Ravi and R. Prithviraj, "A face expression recognition using cnn and lbp," *IEEE 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020.

[115] H. Alshamsi, V. Kepuska, and H. Meng, "Real time automated facial expression recognition app development on smart phones," *8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 384–392, 2017.

[116] M. R. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and R.-t. Roussos, "Facial expression recognition in the wild by disentangling 3d expression from identity," *arXiv*, type = preprint, archivePrefix = arXiv, eprint = 2005.05509, Tech. Rep., 2020.

[117] R. Melaugh, N. Siddique, S. Coleman, and P. Yogarajah, "Facial expression recognition on partial facial sections," in *11th International Symposium on Image and Signal Processing and Analysis*. p, 2019, pp. 193–197.

[118] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern recognition*, vol. 42, no. 3, pp. 425–436, 2009.

[119] L. Chen, W. Lu, J. Ni, W. Sun, and J. Huang, "Region duplication detection based on harris corner points and step sector statistics," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 244–254, 2013.

[120] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 32–39.

[121] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, vol. 56. Citeseer, 2000.

[122] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 962–969, 1993.

[123] 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*, vol. 1. IEEE, 2001, pp. I–I.

[124] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[125] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[126] T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61 677–61 685, 2018.

[127] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2559–2573, 2013.

[128] M. Z. Uddin, J. Lee, and T.-S. Kim, "An enhanced independent component-based human facial expression recognition from video," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 4, pp. 2216–2224, 2009.

[129] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010.

[130] N. Aifanti and A. Delopoulos, "Linear subspaces for facial expression recognition," *Signal Processing: Image Communication*, vol. 29, no. 1, pp. 177–188, 2014.

[131] L. Zhao, Z. Wang, and G. Zhang, "Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[132] Y. Sun and J. Yu, "Facial expression recognition by fusing gabor and local binary pattern features," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 209–220.

[133] 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.

[134] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, 2014.

[135] P. Ekman, "Facial expression and emotion," *American psychologist*, vol. 48, p. 4, 1993.

[136] M. S. Bartlett *et al.*, "Real time face detection and facial expression recognition: development and applications to human computer interaction," in *2003 Conference on computer vision and pattern recognition workshop*. Vol. 5. IEEE, 2003.

[137] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee, 2005.

[138] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[139] S. Berretti *et al.*, "3d facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.

[140] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, "Survey on face expression recognition using cnn," in *2019 5th international conference on advanced computing & communication systems (ICACCS)*. IEEE, 2019, pp. 102–106.

[141] O. M. Parkhi, A. Vedaldi, and A. Zisserman, $vgg/software/vgg_face/$, University of Oxford, 2015.

[142] L. Wolf, T. Hassner, and I. . Maoz, "June)," *Face recognition in unconstrained videos with matched background similarity*, pp. 529–534, 2011.

[143] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013.

[144] ——, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[145] G. Zhao *et al.*, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[146] Y. Chen *et al.*, "Three-stream convolutional neural network with squeeze-and-excitation block for near-infrared facial expression recognition," *Electronics*, vol. 8, p. 4, 2019.

[147] H. Jung *et al.*, "Deep temporal appearance-geometry network for facial expression recognition," arXiv, preprint, 2015.

[148] L. A. Jeni, H. Hashimoto, and T. Kubota, "Robust facial expression recognition using near infrared cameras," *J. Adv*, vol. 16, no. 2, pp. 341–348, 2012.