PROJECT REPORT

ON

HUMAN EMOTION RECOGNITION USING BIASED DEEP LEARNING MODEL

Submitted in complete fulfilment of the requirements

for the award of the degree of

BACHELOR’S IN TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

K. Abhinandu Reddy	A. Ramanand Chowdary
2017BCSE054	2017BCSE056

Under the supervision of

Dr. Ranjeet Kumar Rout

Department of Computer Science and Engineering

National Institute of Technology, Srinagar, J&K

Department of Computer Science and Engineering

National Institute of Technology, Srinagar, J&K

CERTIFICATE

This is to certify that the project titled **HUMAN EMOTION RECOGNITION USING BIASED DEEP LEARNING MODEL** has been completed by **K.Abhinandu Reddy** (2017BCSE054) and **A.Ramanand Chowdary** (2017BCSE056) under my supervision to complete fulfilment of the requirements for the award of the degree of Bachelor’s in Technology in Computer Science and Engineering. It is also certified that the project has not been submitted or produced for the award of any other degree.

Dr. Ranjeet Kumar Rout

Assistant Professor

Department of Computer Science and Engineering

NIT, Srinagar

STUDENTS DECLARATION

We, hereby declare that the work, which is being presented in the project entitled **HUMAN EMOTION RECOGNITION USING BIASED DEEP LEARNING MODEL** in complete fulfilment of the requirements for the award of the degree of Bachelors of Technology in Computer Science and Engineering in the session 2021, is an authentic record of our own work carried out under the supervision of Dr. Ranjeet Kumar Rout, Department of Computer Science and Engineering, National Institute of Technology, Srinagar. The matter embodied in this project has not been submitted by us for the award of any other degree.

Dated: 08-06-2021

Name: K.Abhinandu Reddy

Signature:

Name: A.Ramanand Chowdary

Signature:

ACKNOWLEDGEMENT

It is our privilege to express our sincere thanks to our project mentor, Dr. Ranjeet Kumar Rout, for his valuable inputs, able guidance, encouragement, whole–hearted cooperation, and constructive criticism throughout the duration of our project on the topic "HUMAN EMOTION RECOGNITION USING BIASED DEEP LEARNING MODEL".

We take this opportunity to express gratitude to all of the Department faculty members for their help and support. We would like to thank our parents for their unceasing encouragement, support, and attention. We are also grateful to each other, to support the group throughout this venture.

We also place on record our sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

K.Abhinandu Reddy

2017BCSE054

A.Ramanand Chowdary

2017BCSE056

8th Semester

BTech, Computer Science and Engineering

NIT Srinagar

Abstract

The focus of this study is on computer automated perception of human emotion. The ability of robots to recognize emotions will be vital for the development of service robotics in the coming years. However, we have not achieved the recognition of certain emotions in a sufficiently effective way yet. Therefore, we have not introduced in the market products using it. If emotion recognition is a very challenging task even for people, it is even more for robots, since they do not feel any emotion and can not empathize.

We build and test a compound hierarchical system that attempts to interpret human emotion in real time using face detection and tracking algorithms in conjunction with our facial expression analysis methodology.

We will be using KDFE and SFEW dataset in this report for all evaluations and novel deep neural networks and related training strategies that are designed for FER based on both static images and dynamic image sequences, and discuss their advantages and limitations.

Finally, we review the remaining challenges and corresponding opportunities in this field as well as future directions for the design of robust FER systems.

TABLE OF CONTENTS PageNo

1. 1.1.Motivation …………………………………………………………..1

1.2. Necessity ……………………………………………….............2

2. Facial Expression Database and Evaluation……………...................3

2.1 KDEF …………………………………………………………..4

2.2 SFEW …………………………………………………………..5

3. Related Work ………………………………………………………...5

4. Facial Expression Recognition ………………………………...........6

4.1 Pre-Processing ………………………………………...........11

4.2 Face Alignment ……………………………………………….11

4.3 Data Augmentation.……………………………………..........12

4.4 Face Normalization.……………………………………………....12

4.5 Illumination Normalization………………………………....12

4.6 Pose Normalization ………………………………………...13

5. Image Pre-Processing ………………………………………………13

5.1 Converting color to Grayscale…………………………...14

5.2 Image Cropping ……………………………………………14

5.3 Down Sampling ……………………………………………15

6. Design and CNN Structure Description ……………………………16

7.Deep Neural Network ………………………………………………17

8. Results and Proposed Models ……………………………….....20

8.1 Progressive Resizing ……………………………………………20

8.2 Resnet50+customNet ……………………………………………23

8.3 Our State of Art Model ……………………………………......25

9. Future Work ……………………………………………………...29

10. References …………………………………..………………30

TABLE OF FIGURES

1: Hierarchy of datasets …………………………………………………….. 3

2: KDEF Facial Expression Dataset ……………………………….……….. 4

3: SFEW ……………………..………………………………….………. 5

4: Facial Expression Prediction Process ……………….…………….………13

5: Grayscale Conversion ……………………………………….………. 14

6: Haar Cascade ………………………………………………..………. 15

7: Down Sampling ……………………..……………………………. 16

8: Process ……………………..…………………………..………. 17

9: AlexNet ……………………..…………………………….………. 18

10: VGGNet ………..…………………………………………..………. 18

11: ResNet ………...……………………………………….………19

12: Google Net ………..…………………………………..………. 19

13: Progressive Resizing ………………………………………….………..20

14: Graph Epoch vs Accuracy and Epochs vs Loss on KDEF ……………. 21

15: Graph Epoch vs Accuracy and Epochs vs Loss on SFEW ……..…….. 22

16: Graph Epoch vs Accuracy and Epochs vs Loss on KDEF of Resnet50. 24

17: CNN Architecture Block ……………………………..…………….. 25

18: Proposed model structure ……………………………...……………. 26

19: Complete CNN architecture ………………...………………………27

20: Classification Report …………………….……………………………..28

21: Graph Epoch vs Accuracy and Epochs vs Loss of Our Process …...…..28

1.1 MOTIVATION

Facial expression, which is a fundamental mode of transporting human's emotions, plays a significant role in our daily communication. Facial expression recognition is a complex and interesting problem, and finds its applications in driver safety, health-care, human–computer interaction etc. Due to its wide range of applications, facial expression recognition has received substantial attention among the researchers in the area of computer vision [1,2,3].

Emotion recognition is a natural capability in human beings. However, if we are to ever create a humanoid robot that can interact and emote with its human companions, the difficult task of emotion recognition will have to be solved. The ability for a computer to recognize human emotion has many highly valuable real world applications. Consider the domain of therapy robots which are designed to provide care and comfort for infirm and disabled individuals. These machines could leverage information on a patient's current and evolving state of mind, in order to tailor personalized strategies for patient care and interaction. For example, when a patient is upset or unhappy, a more effective strategy may be to take a moment to recognize the emotion and offer sympathy. The patient, feeling heard and validated, may be more likely to be cooperative in subsequent requests issued by the machine.

This thesis tests some state-of-the-art machine learning and deep learning models for emotion recognition using 2 different datasets.

Artificial intelligence has had increasing involvement in any scope of human life. The technologies are adapted to the needs of the human being and artificial intelligence is what makes this adaptation between technology and humans possible.

As early as the twentieth century, Ekman and Friesen [4] defined six basic emotions based on cross-culture study [5], which indicated that humans perceive certain basic emotions in the same way regardless of culture. These prototypical facial expressions are anger, disgust, fear, happiness, sadness, and surprise. Contempt was subsequently added as one of the basic emotions [6]. Recently, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal [7].

1.2 Necessity

FER systems can be divided into two main categories according to the feature representations: static image FER and dynamic sequence FER. In static-based methods the feature representation is encoded with only spatial information from the current single image, whereas dynamic-based methods consider the temporal relation among contiguous frames in the input facial expression sequence.

Although very deep or wide networks-based FER approaches usually perform reasonably well, they still have a few problems related to processing time and memory consumption, which are associated with the multitudinous parameters in the training and inference processes. However, in most embedded systems, including intelligent vehicle systems, real-time processing of DNNs is a heavy burden.

The majority of the traditional methods have used handcrafted features or shallow learning (e.g., local binary patterns (LBP) [8], LBP on three orthogonal planes (LBP-TOP) [9], non-negative matrix factorization (NMF) [10] and sparse learning [11]) for FER.

Recent developments in deep learning reduce the burden of handcrafting the features. Deep learning approaches perform well for all the above-mentioned tasks by learning an end-to-end mapping from the input data to the output classes. Out of all the learning based techniques, convolutional neural network (CNN) based techniques are preferred, where the extracted features are combined using a dense layer and the expressions are classified based on the output score of the soft-max layer.

2. FACIAL EXPRESSION DATABASES AND EVALUTION

Having sufficient labelled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system. In this section, we discuss the publicly available databases that contain basic expressions and that are widely used in our reviewed papers for deep learning algorithms evaluation. We also introduce newly released databases that contain a large amount of affective images collected from the real world to benefit the training of deep neural networks.

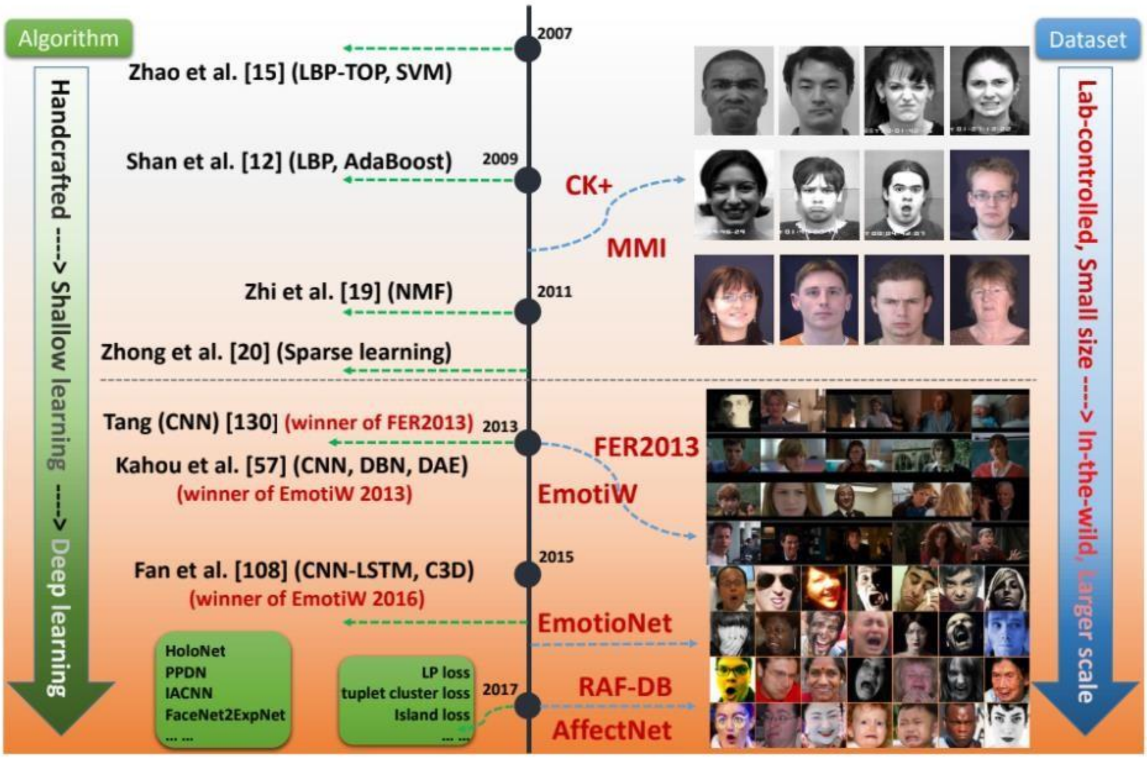

Algorithm
Dataset
2007
Zhao et al. [15] (LBP-TOP, SVM)
Shan et al. [12] (LBP, AdaBoost)
2009
CK+
MMI
Zhi et al. [19] (NMF)
2011
Zhong et al. [20] (Sparse learning)
Tang (CNN) [130] (winner of FER2013)
2013
FER2013
Kahou et al. [57] (CNN, DBN, DAE)
(winner of EmotiW 2013)
EmotiW
2015
Fan et al. [108] (CNN-LSTM, C3D)
(winner of EmotiW 2016)
EmotioNet
HoloNet
PPDN
IACNN
FaceNet2ExpNet
... ...
LP loss
tuplet cluster loss
Island loss
... ...
2017
RAF-DB
AffectNet
Handcrafted ----> Shallow learning ----> Deep learning
Lab-controlled, Small size ----> In-the-wild, Larger scale

Fig 1: Hierarchy of datasets

We are interested in two datasets KDEF and SFEW

2.1 Karolinska Directed Emotional Faces [KDEF]

Karolinska Directed Emotional Faces was collected by scientists Ellen Goeleven, Rudi De Raedt, Lemke Leyman and Bruno Verschuere from Ghent University in Belgium [11]. It contains 4900 JPEG images showing 70 people (35 women and 35 men) displaying seven different facial expressions. Each expression is recorded from 5 different angles. All the participants were amateur actors between 20 and 30 years old. For the election of the individuals, mustaches, beards, eyeglasses, earrings, and make-up were exclusion criteria. Some samples can be seen in figure 1. We have use only Female images for training and male images for testing.

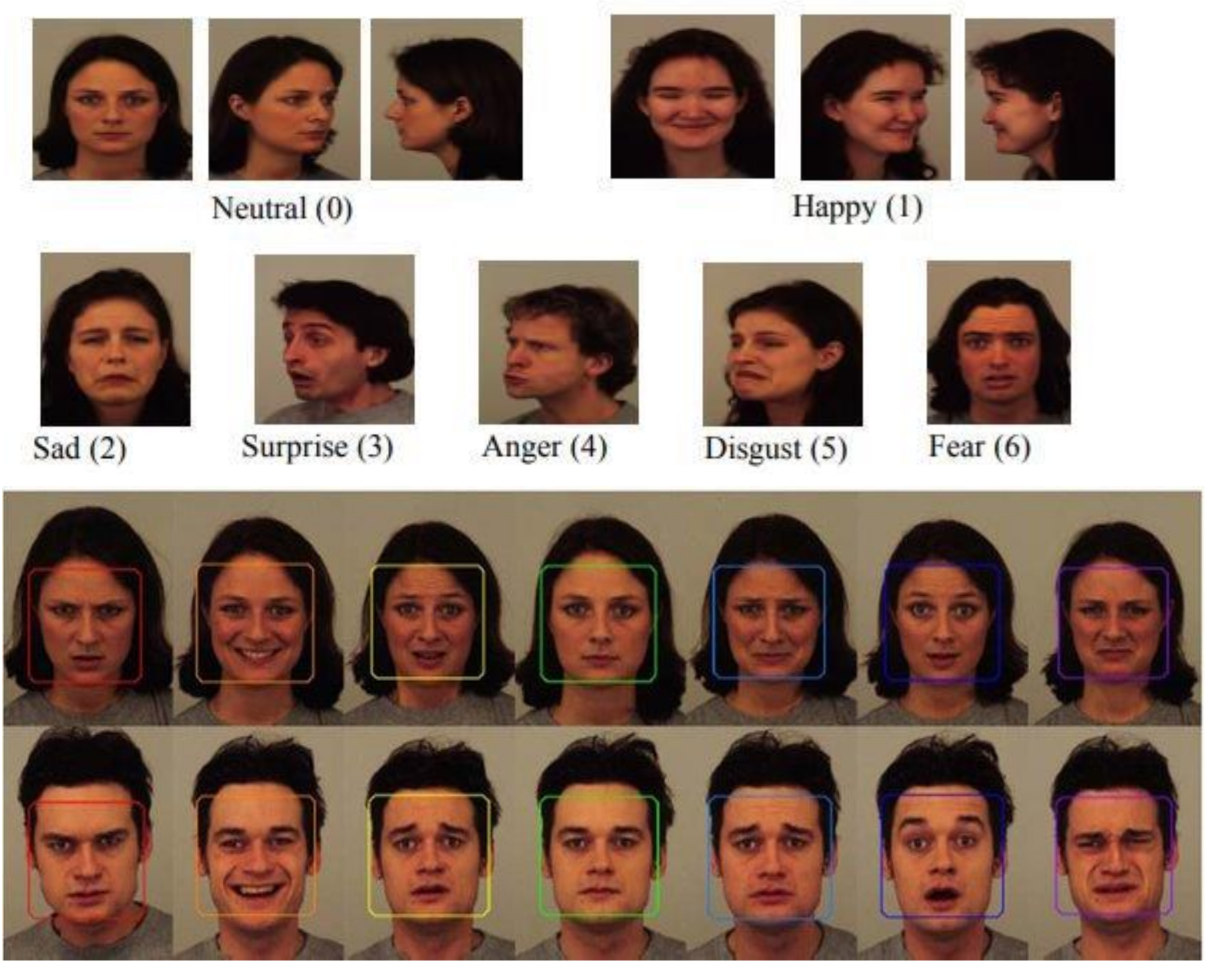


Fig 2: KDEF Facial Expression Dataset.

2.2 SFEW :

“Static Facial Expressions in the Wild (SFEW) Database” [12], selected from frames of a temporal facial expressions database, Acted Facial Expressions In the Wild (AFEW). This database includes unconstrained facial emotions, different head poses, wide age scale, occlusions, various focus and near to real-world illumination. AFEW was derived from movies clips. While movies are usually shot in slightly controlled circumstances, they provide close to real-life environments that are more realistic than other current datasets that were recorded in lab environments. SFEW has both frontal and non-frontal faces and different illumination conditions (Figure 2. shows some sample images from the database) which are very like real-world scenarios. SFEW images are labelled into basic expression angry, disgust, fear, happy, sad, surprise and the neutral class. Labels of the training and validation sets are publicly available, whereas those of the testing set are held back by the challenge organizer.

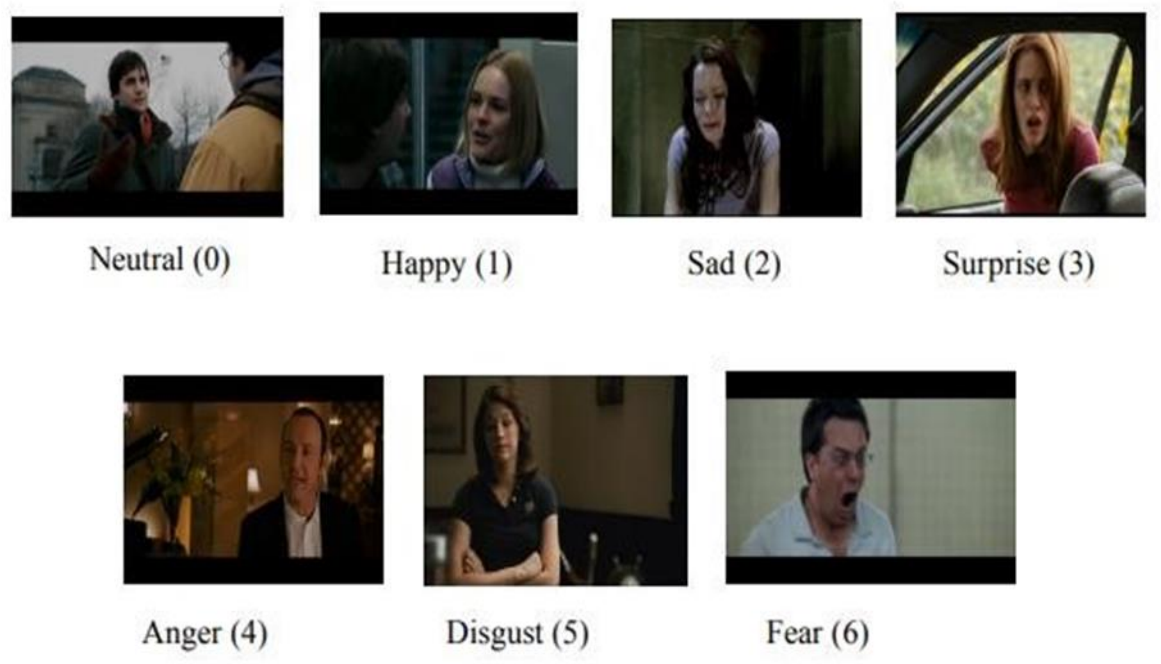


Fig 3: SFEW

3. RELATED WORKS

Facial expression recognition is an example of a problem that humans are good at solving, but computers are not [13]. A lot of work has been invested in trying to have computers reach the same accuracy as humans, and some examples of these attempts are highlighted here. Several emotion recognition methods have been developed in the last few years, and a lot of advancement has been achieved in this area [14]. This section focuses on some closely related approaches to our work.

Ma and Khorasani proposed two facial expression recognition techniques [15]. They used Lower-frequency 2-D DCT coefficients of binarized edge images as features for recognition. A constructive one-hidden-layer (OHL) feedforward neural network (OHL-NN) was used for the first approach, and K-means algorithm as classifiers was used in the second approach. Their approach significantly reduced the size of the neural network by pruning technique while improving the generalization capability and the recognition rate. The technique they proposed was tested to a database with images of 60 men, all having five facial expressions (neutral, smile, anger, sadness, and surprise). They used 40 images for network training, and the remaining 20 for generalization and testing. Confusion matrices were used for performance evaluation. They achieved high recognition rates, 100% for the training, and 93.75% for generalizing images. Even though they achieved high accuracy, they only had five distinct emotions.

Facial emotion recognition is usually accomplished sequentially in three individual stages: feature (shape of the eyes, nose, cheekbones, and jaw) extraction and learning, feature selection, and classifier modelling [16]. “Extensive empirical researches are required to search for an optimal combination of feature representation (feature extraction), feature set (feature selection), and classifier to achieve good recognition performance” [17]. P. Liu, et al. [17] presented a new Boosted Deep Belief Network (BDBN) for implementing the three training stages iteratively in a unified loopy framework: “BDBN framework consists of two interconnected learning processes: a bottom-up unsupervised feature learning (BU-UFL) method that learns hierarchical feature representations given input data and a boosted top-down supervised feature strengthen (BTD- SFS) process that refines the features jointly in a supervised manner.” This method used input image patches instead of the entire image for facial emotion recognition. A collection of features, which is useful to distinguish expression-related facial appearance/shape changes, could be detected and selected to develop a boosted powerful classifier in a statistical process through this BDBN framework. According to the authors, through this framework, strong classifiers improved iteratively as learning continued and discriminative capabilities of selected features were strengthened according to their comparative importance to the strong classifier via a joint fine-tune process [17]. Their experiments were conducted using two public databases of static images, Extended CohnKanade (CK+) database [18] and JAFFE [19] database, and achieved an accuracy of 96.7% and 68.0% respectively. They focused on six basic expressions and took about eight days to complete the overall training.

C. Shan et al. used Local Binary Patterns (LBP), which is not using a neural network, as the feature extractor [20]. In this paper, they merged and compared different machine learning techniques such as Support Vector Machine, Linear Discriminant Analysis, etc. for facial expressions. They empirically assessed facial representation based on LBP and based on statistically local features for person-independent facial expression recognition. Through their experiments, they observed that LBP features operated stably and robustly across a useful range of low resolutions of face images and produced promising performance in compressed low resolution video sequences captured in real-world circumstances. Using SVM and LBP, they achieved 95.1% accuracy in the Cohn-Kanade [18] database and used a 10-fold cross-validation scheme for performance evaluation.

S. Borah and S. Konwar [21] presented an artificial neural network (ANN)-based human facial expression recognition method. The authors detected 22 facial feature points through an automatic technique and generated feature vector by calculating the Euclidian distance between certain points. These vectors were given as input to a multi-layer perceptron (MLP) neural network. They used connected component analysis for face detection and 2D Color Space Skin Clustering method for skin color detection. In this paper, the facial expressions were classified into seven basic expressions (anger, disgust, sad, surprise, fear, happy, and neutral) using a feed forward back propagation neural network. The best result achieved in their work was an accuracy rate of 85% in the color FERET database. Even though the method proposed in this paper could automatically locate facial feature points at high accuracy for most front faced images, it was limited when angle rotated images were involved. They did not mention anything about training time.

An approach using CNN for feature and classification has also been used before [22]. The authors proposed a simple model for facial emotion recognition which uses a combination of CNN and specific image pre-processing steps that achieve comparatively higher accuracy. CNN and most of the machine learning methods helped produce high accuracy depending on a given feature set. The CNN models used raw images as input rather than hand-coded features. A typical architecture of a CNN had sequenced layers including an input layer, convolutional layers (pooling layer + rectified linear units), dense layers (multilayer fully-connected networks), and an output layer. The CNNs achieved better results than traditional neural networks. A. T. Lopes et al. [22] used a single CNN for performing three learning stages such as feature learning, feature selection, and classification of their work. The CNN was trained using many input images and their emotions. The weights of neural networks in each layer were adjusted while training. For recognizing an unknown image, the system applied a sequence of image pre-processing steps including spatial normalization, image cropping, downsampling, and intensity normalization on test images. They performed their experiments using Extended Cohn-Kanade (CK+) database [23] and achieved 97.81% accuracy. We are using a few of their image pre-processing steps in our implementation. Breuer et al. [24] demonstrate the capability of the CNN network trained on various FER datasets by visualizing the feature maps of the trained model, and their corresponding FACS action unit. Motivated by Xception architecture proposed in [25], Arriaga et al. [20] proposed mini-Xception. Jung et al [26] proposed two different deep network models for facial expression recognition. The first network extracts temporal appearance features, whereas the second extracts temporal geometric features and these networks are combined and fine tuned in the best possible way to obtain better accuracy from the model. Motivated by these two techniques, we have trained and obtained multiple models, the details of which are explained

For the task of facial emotion recognition, the current state of the art model [27] proposed a miniature version of VGG net, called VGG13. The network has 8.75 million parameters. The dataset used is the FERplus dataset, which has 8 classes, adding neutral to the existing seven classes. The reported test accuracy is ≈ 84%. In 2014, G. Levi et al. [28] improved emotion recogition using CNN. They convert images to local binary patterns. These patterns are mapped to a 3D metric space and used as an input to the existing CNN architectures, thus addressing the problem of appearance variation due to illumination. They trained the existing VGG network [29], on CASIA Webface dataset, and then used transfer learning to train the Static Facial Expressions in the Wild (SFEW), to address the problem of small size of SFEW dataset

Architecutre Details	**Accuracy %**
Original RTNN	83.16
STL +RTNN	84.08
RNN + Laplacian RTNN	84.39
STL with RTNN + Gradient RTNN	85.10
STLwith RTNN+ Laplacian RTNN	85.51
STL with original gradient and Laplacian RTNN	88.16

Table:1 RTNN methods Accuracy

M. Shin et al. recommended a baseline CNN structure and image pre-processing methodology to improve facial expression recognition [30]. They experimented with four different CNN structures and five types of pre processed images (raw, histogram equalization, isotropic smoothing, diffusion-based normalization, and difference of Gaussian) and suggested an efficient baseline CNN structure and pre-processing for facial expression recognition. The authors have used five different datasets: JAFFE, FER-2013, SFEW2.0, CK+ (extended CohnKanade), and KDEF (Karolinska Directed Emotional Faces) [31] for performance evaluation. All datasets used in their work included seven types of the same emotional expression and were fully labeled. Their experiment result showed that three-layer network with a simple convolutional and a max pooling layer with histogram equalization images performed best among the four networks they used. Their highest accuracy rates were 50.61% with JAFFE and 59.15% with KDEF. They did not mention anything about the time taken to train their networks. We are also using all four of these networks in our experiment and this paper as our baseline.

So far, a lot of works for facial expression recognition using CNN-based deep neural networks were introduced, and their CNN structures and image pre-processing methods were all different. The main reason for the differences is that the selection of pre-processing methods and CNN structures were not main focuses in their research. However, proper selection of CNN structure and appropriate image pre-processing of input data are very important for improving results. In this paper, we are using four different network structures and several image preprocessing steps on different public databases to compare performance.

4. FACIAL EXPRESSION RECOGNITION

In this section, we describe the three main steps that are common in FER, i.e., pre-processing, deep feature learning and deep feature classification.

4.1 Pre-processing

Variations that are irrelevant to facial expressions, such as different backgrounds, illuminations and head poses, are fairly common in unconstrained scenarios. Therefore, before training the deep neural network to learn meaningful features, pre-processing is required to align and normalize the visual semantic information conveyed by the face.

4.2 Face alignment

Face alignment is a traditional pre-processing step in many face-related recognition tasks. We list some well-known approaches and publicly available implementations that are widely used in deep FER. Given a series of training data, the first step is to detect the face and then to remove background and nonface areas. The Viola-Jones (V&J) face detector is a classic and widely employed implementation for face detection, which is robust and computationally simple for detecting near-frontal faces.

4.3 Data augmentation

Deep neural networks require sufficient training data to ensure generalizability to a given recognition task. However, most publicly available databases for FER do not have a sufficient quantity of images for training. Therefore, data augmentation is a vital step for deep FER. Data augmentation techniques can be divided into two groups: on-the-fly data augmentation and offline data augmentation.

4.4 Face normalization

Variations in illumination and head poses can introduce large changes in images and hence impair the FER performance. Therefore, we introduce two typical face normalization methods to ameliorate these variations: illumination normalization and pose normalization (frontalization).

4.5 Illumination normalization:

Illumination and contrast can vary in different images even from the same person with the same expression, especially in unconstrained environments, which can result in large intra-class variances. In [60], several frequently used illumination normalization algorithms, namely, isotropic diffusion (IS)-based normalization, discrete cosine transform (DCT)-based normalization [85] and difference of Gaussian (DoG), were evaluated for illumination normalization.

4.6 Pose normalization:

Considerable pose variation is another common and intractable problem in unconstrained settings. Some studies have employed pose normalization techniques to yield frontal facial views for FER, among which the most popular was proposed by Hassner. Specifically, after localizing facial landmarks, a 3D texture reference model generic to all faces is generated to efficiently estimate visible facial components. Then, the initial frontalized face is synthesized by backprojecting each input face image to the reference coordinate system.

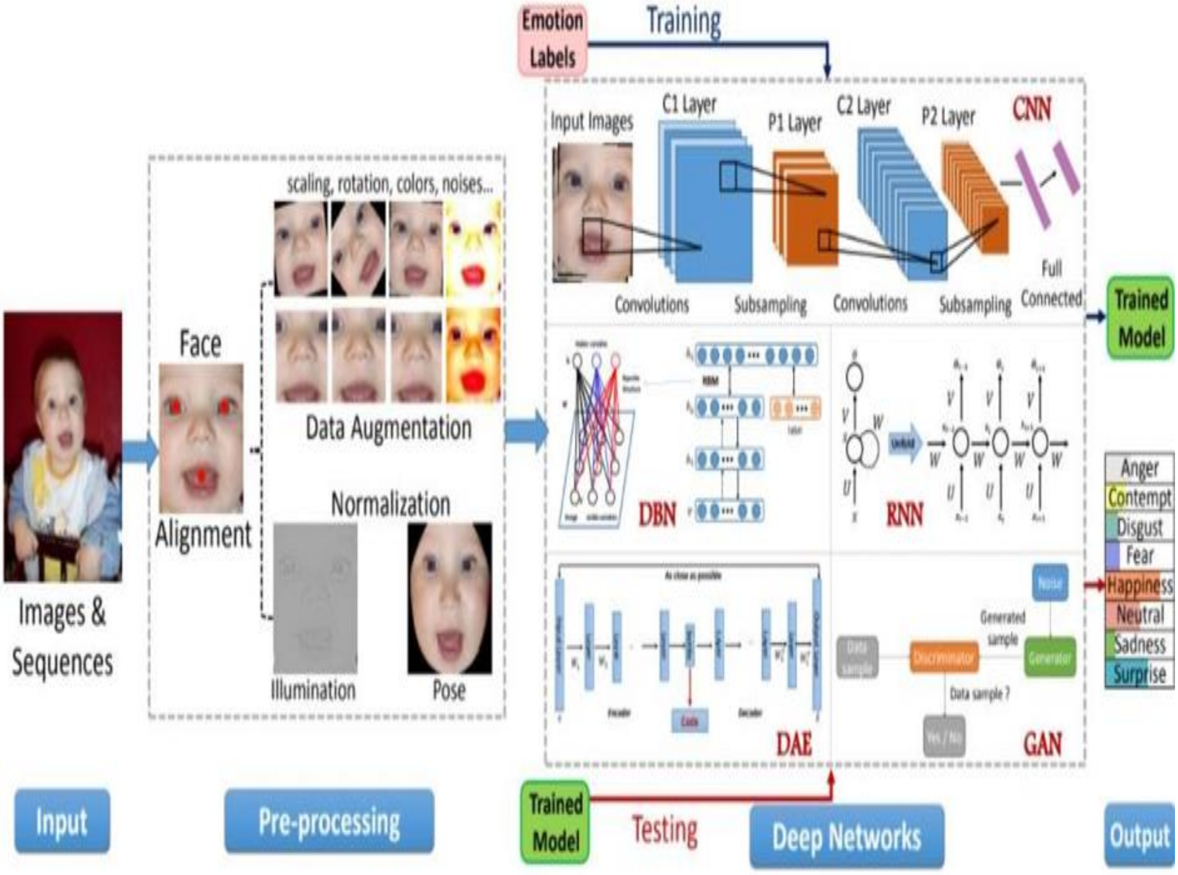


Fig 4: Facial Expression Prediction Process

5. Image Pre-processing

Our proposed method of facial recognition consists of two steps: image preprocessing and performance evaluation. Performance evaluation comprises of two phases: training and testing. Quality and performance of our neural network model are constrained by the quality of data we consider for testing and training. All existing deep neural networks are susceptible to the quality of images such as distortions, particularly blurriness, rotations, brightness, and image size.

5.1 Converting Color to Grayscale:

Processing color images are complex and time consuming, and grayscale provides an easy way out. For example, when we use color images, all operations should perform on all three-image planes (R, G, and B), while grayscale has only one image plane. As shown in Figure 6, the original image is the color image, and the color information doesn't benefit us to recognize important edges or other characteristics that are important for expression recognition. To reduce the complexity of the code, we have converted color images to grayscale. Another important advantage of converting color to gray is that we can reduce the processing time. While grayscale conversion can cause loss of color information that is required in many image processing applications, color info does not add any value in our experiment.

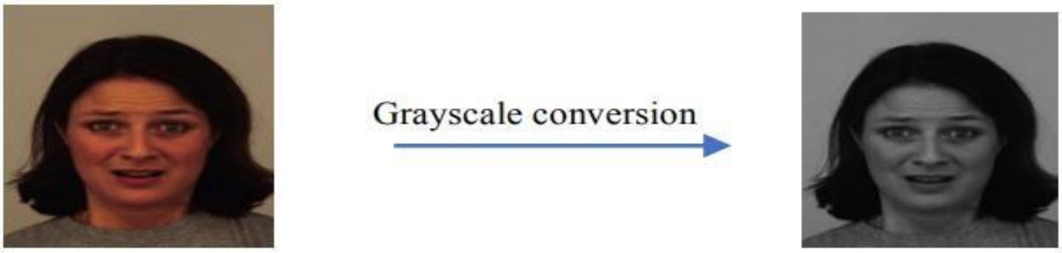


Fig 5: Grayscale Conversion

5.2 Image Cropping:

Most of the papers on expression recognition used the face cropping technique to achieve high accuracy. As shown in Fig. 6, the image from our dataset has a lot of background information that does not give any important information about the expression to the expression recognition procedure. Moreover, this extra background information could decrease the accuracy rate of the recognition of expression because the neural network model has an extra problem to solve, differentiating between foreground and background information. We have used the Viola-Jones object detection framework for detecting faces and cropped face parts from the image.

In this method, the object is detected using Haar feature-based cascade classifiers, and this method was proposed by Paul Viola and Michael Jones. After the cropping process, all parts of the image that do not add any value to the expression classification procedure are removed. Thus, the resulted image does not contain facial parts that do not contribute to the expression (e.g., hair, ears, etc.). This step helped us to extract meaningful features from the image for classification. Image cropping example is shown in Figure

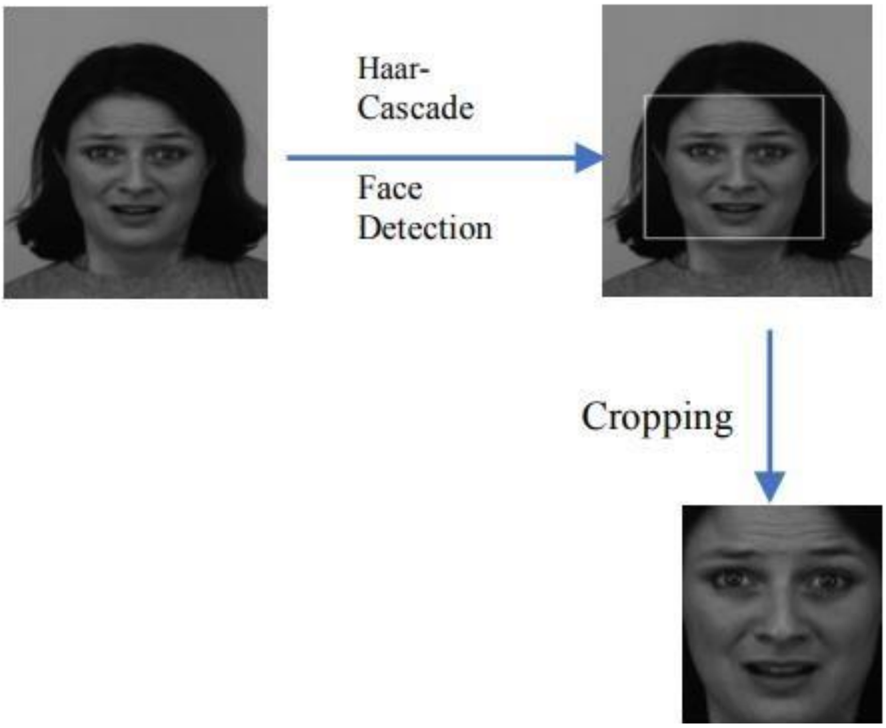


Fig 6: Haar Cascade

5.3 Down Sampling:

Downsampling, downscaling images with geometric transformations without losing the quality of the image helps us to reduce the amount of data we have. As the size of the image data increases, a lot of memory is required for the neural network to perform their operations. Nowadays machines we can use for image processing applications have limited memory.

.

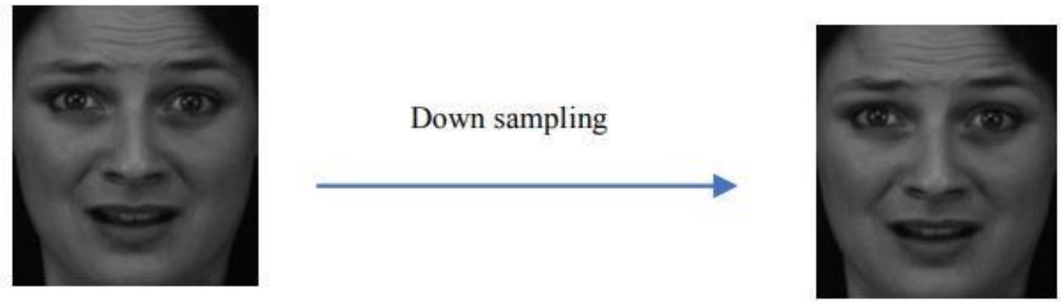


Fig 7: Down Sampling

We applied downsampling to the images without losing the important information to avoid the problem of memory consumption. The final images from all datasets are 256 * 256 pixels. An example of downsampling is shown Figure

6. Design and CNN Structure Description

Our design approach for recognizing emotions from an image is shown in Figure 10. Initially, we have sets of images. This image data is prepared using specific image processing steps such as grayscale conversion, cropping, and downsampling. After data preparation, resulting images will be of good quality and quantity. These prepared images are used for feature extraction and classification. These two processes are done using CNN. This is the critical step in facial expression recognition. Once classification is done, performance is evaluated using 5-fold cross-validation.

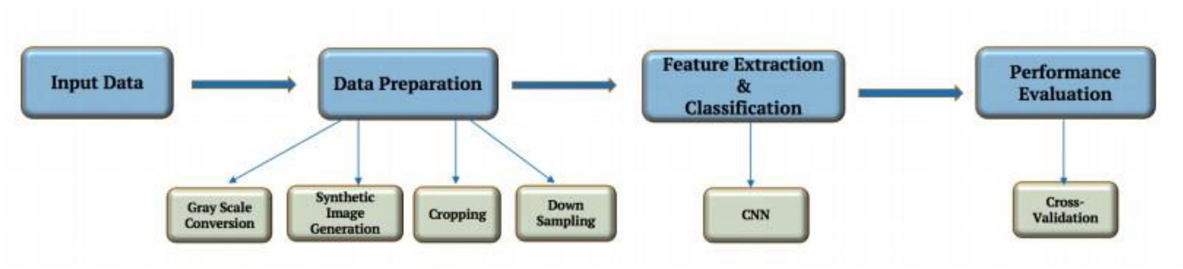


Fig 8: Implementation Process

7. Deep neural networks

The traditional architectures of these deep neural networks FOR FEATURE LEARNING ARE:

7.1. Convolutional neural network (CNN):

*A CNN consists of three types of heterogeneous layers: convolutional layers, pooling layers, and fully connected layers. The convolutional layer has a set of learnable filters to convolve through the whole input image and produce various specific types of activation feature MAPS.

* The convolution operation is associated with three main benefits: local connectivity, which learns correlations among neighbouring pixels; weight sharing in the same feature map, which greatly reduces the number of the parameters to be learned; and shift-invariance to the location of the object.

Some well-known CNN models that have been applied are AlexNet, VGGNet ,GoogleNet ,ResNet.

AlexNet: The network consists of 5 Convolutional (CONV) layers and 3 Fully Connected (FC) layers. The activation used is the Rectified Linear Unit (ReLU).

The structural details of each layer in the network can be found in the table below

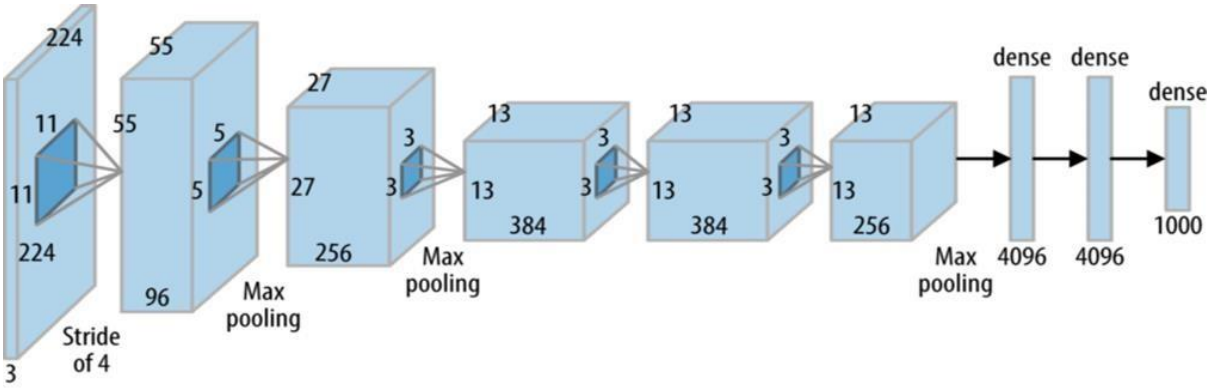


Fig 9: AlexNet

VGGNet:

VGGNet was born out of the need to reduce the # of parameters in the CONV layers and improve on training time.

There are multiple variants of VGGNet (VGG16, VGG19, etc.) which differ only in the total number of layers in the network

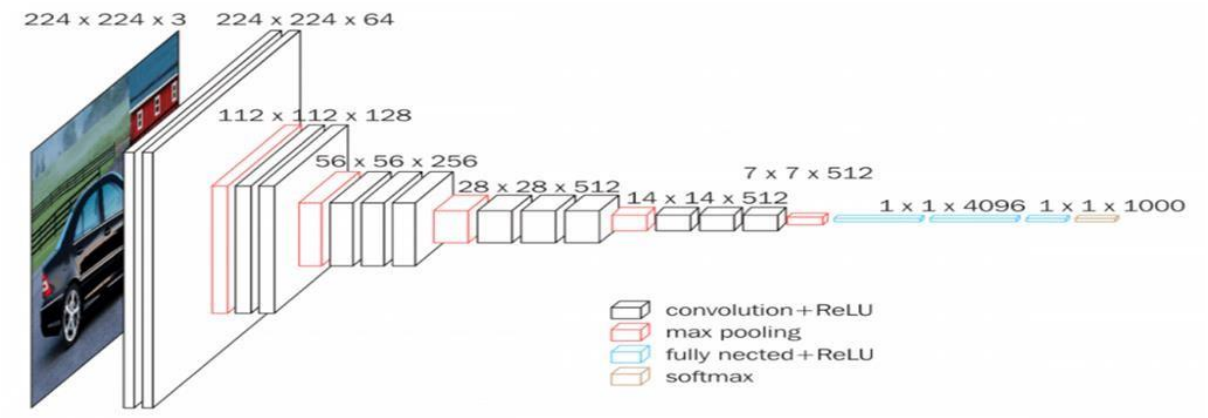


Fig 10: VGGNet

ResNet:

Residual Neural Network is a novel architecture with “skip connections” and features heavy batch normalization.

This technique able to train a NN with 152 layers while still having lower complexity than VGGNet. It achieves a top-5 error rate of 3.57%

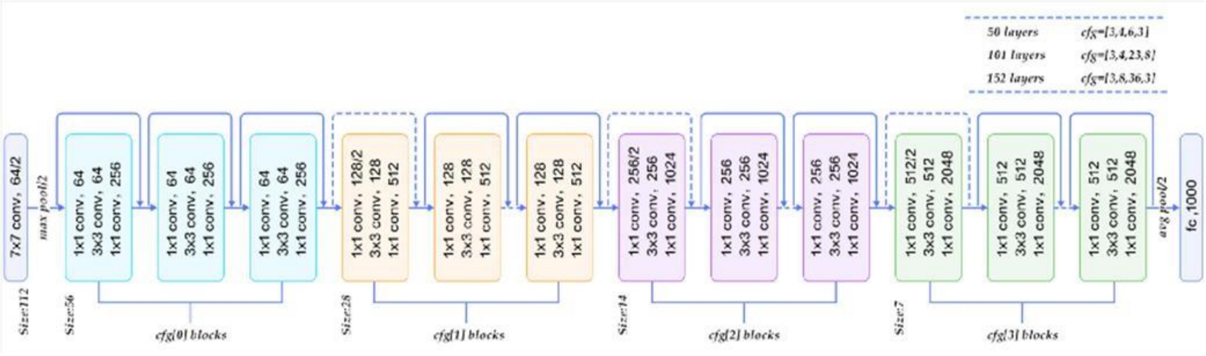


Fig 11: ResNet

Google Net:

GoogLeNet is a convolutional neural network that is 22 layers deep. The GoogLeNet architecture is very different from previous state-of-the-art architectures such as AlexNet and ZF-Net. It uses many different kinds of methods such as 1×1 convolution and global average pooling that enables it to create deeper architecture.

Total Number of operations : (14 x 14 x 48) x (5 x 5 x 480) = 112.9 M With 1×1 convolution :

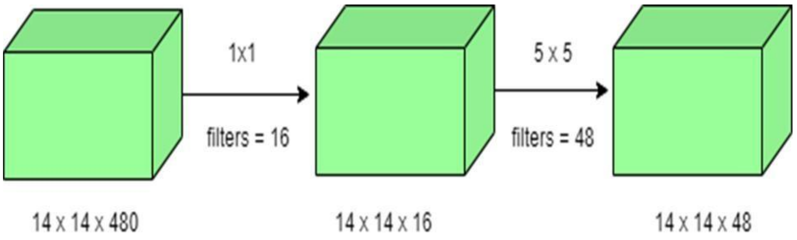


Fig 12: Google Net

8. PROPOSED MODELS

8.1. Progressive resizing

It is the technique to sequentially resize all the images while training the CNNs on smaller to bigger image sizes. Progressive Resizing is described briefly in this terrific fast ai course, “Practical Deep Learning for Coders”. A great way touse this technique is to train a model with smaller image size say 64x64, then use the weights of this model to train another model on images of size 128x128 and so on. Each larger-scale model incorporates the previous smaller scale model layers and weights in its architecture.

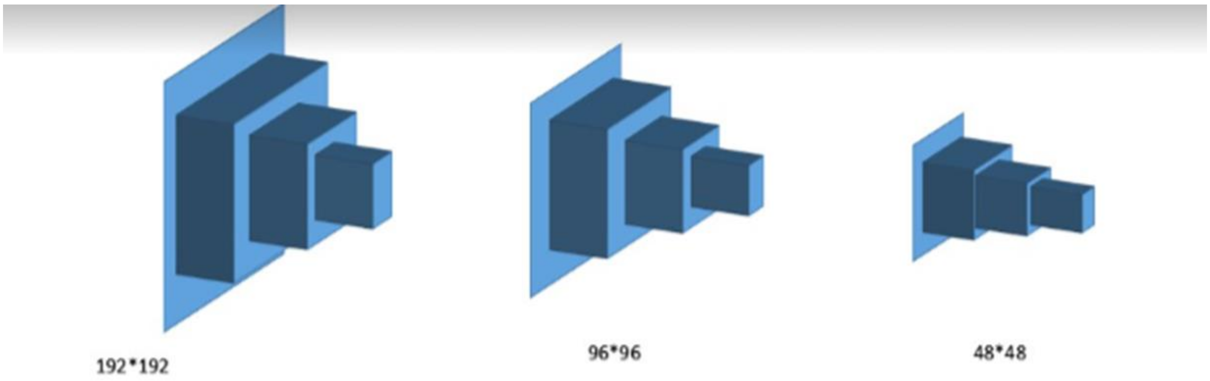
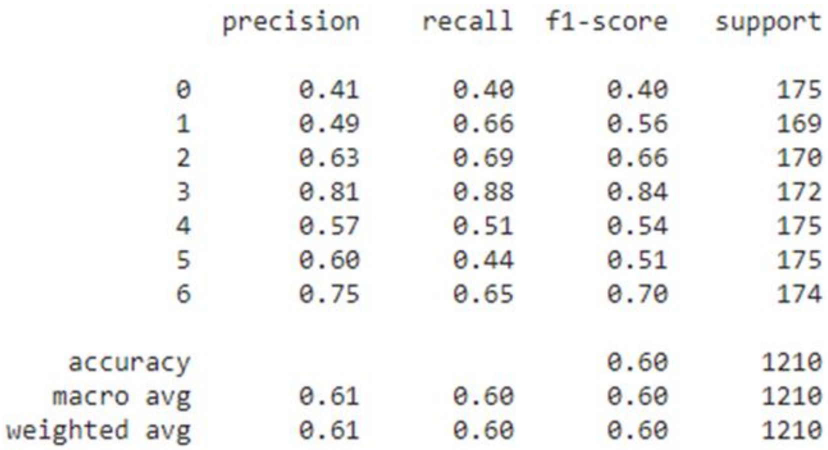


Fig 13: Progressive Resizing

Results

KDEF



```
              precision    recall  f1-score   support

           0       0.41      0.40      0.40       175
           1       0.49      0.66      0.56       169
           2       0.63      0.69      0.66       170
           3       0.81      0.88      0.84       172
           4       0.57      0.51      0.54       175
           5       0.60      0.44      0.51       175
           6       0.75      0.65      0.70       174

    accuracy                           0.60      1210
   macro avg       0.61      0.60      0.60      1210
weighted avg       0.61      0.60      0.60      1210
```

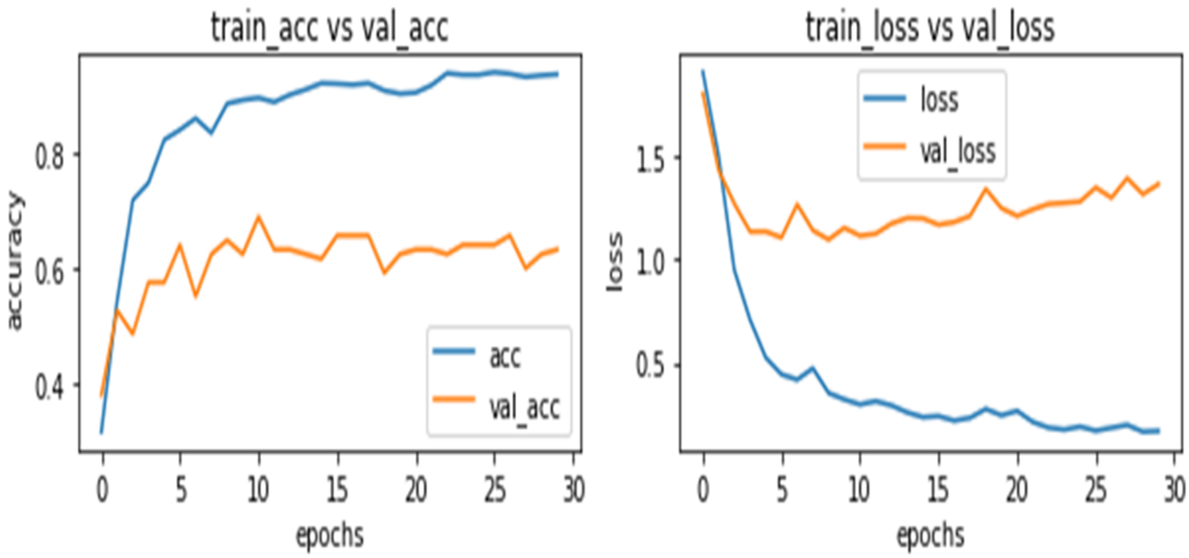
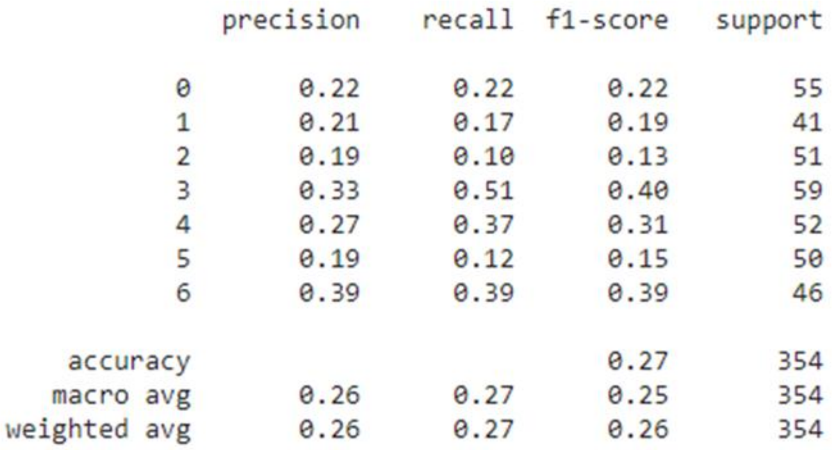


Fig 14: Graph Epoch vs Accuracy and Epochs vs Loss

SFEW



```
              precision    recall  f1-score   support

           0       0.22      0.22      0.22        55
           1       0.21      0.17      0.19        41
           2       0.19      0.10      0.13        51
           3       0.33      0.51      0.40        59
           4       0.27      0.37      0.31        52
           5       0.19      0.12      0.15        50
           6       0.39      0.39      0.39        46

    accuracy                           0.27       354
   macro avg       0.26      0.27      0.25       354
weighted avg       0.26      0.27      0.26       354
```

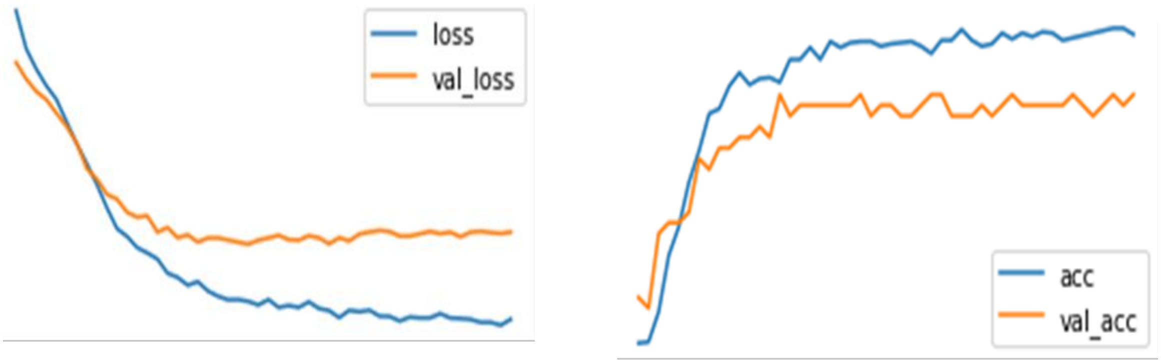
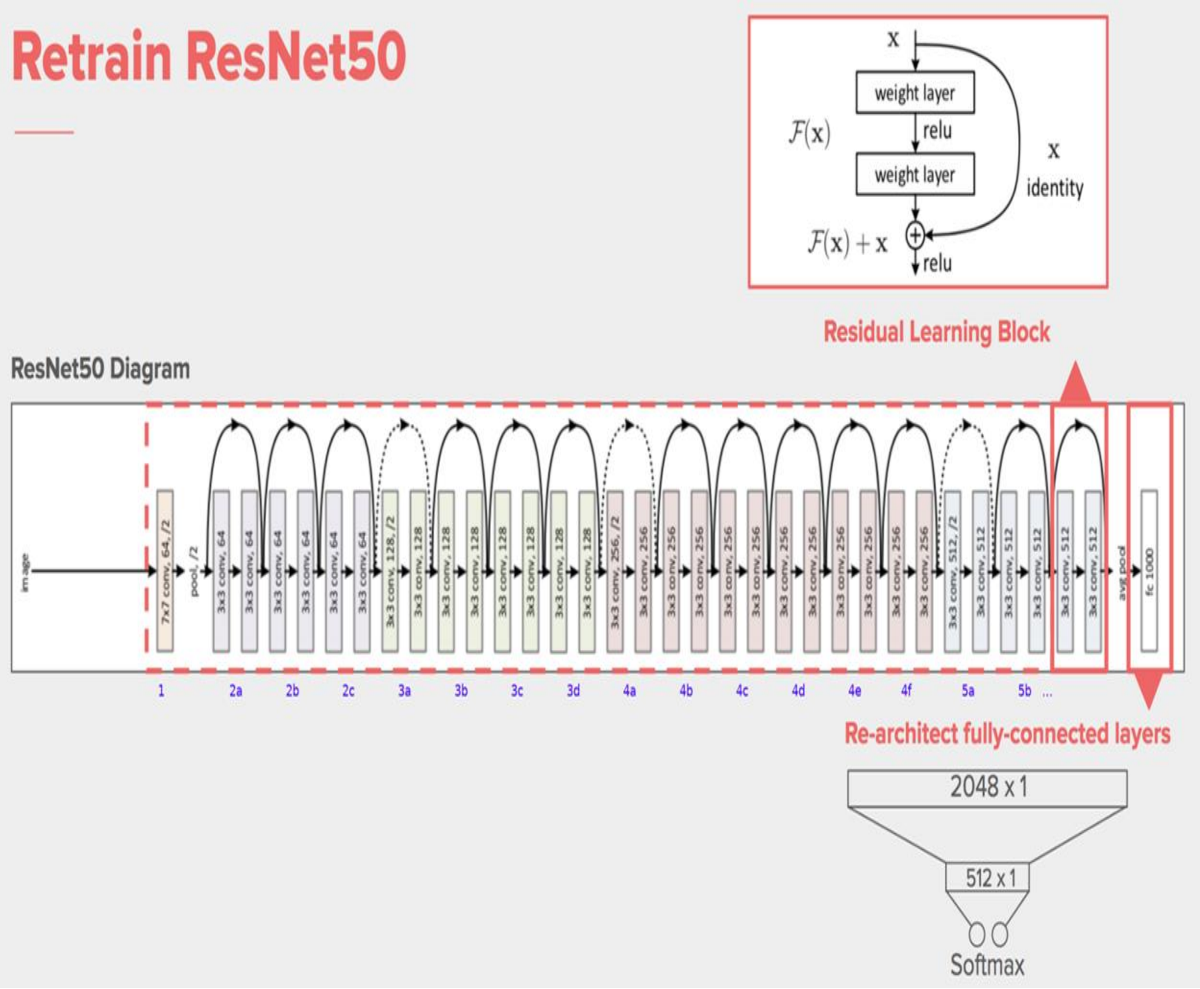


Fig 15: Graph Epoch vs Accuracy and Epochs vs Loss

8.2. ResNet50

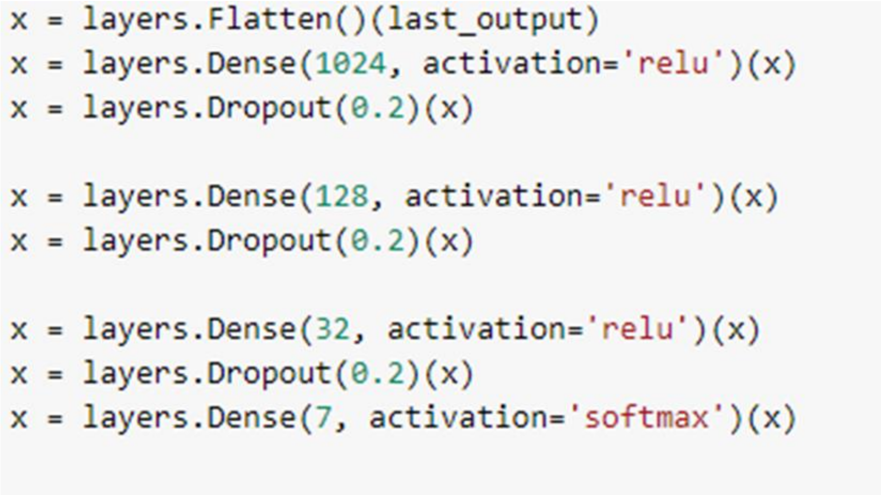
Resnet is short name for Residual Network that supports Residual Learning. The 50 indicates the number of layers that it has. So Resnet50 stands for Residual Network with 50 layers.

It is a widely used ResNet model and has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer



RESULT

KDEF DATASET



```
x = layers.Flatten()(last_output)
x = layers.Dense(1024, activation='relu')(x)
x = layers.Dropout(0.2)(x)

x = layers.Dense(128, activation='relu')(x)
x = layers.Dropout(0.2)(x)

x = layers.Dense(32, activation='relu')(x)
x = layers.Dropout(0.2)(x)
x = layers.Dense(7, activation='softmax')(x)
```

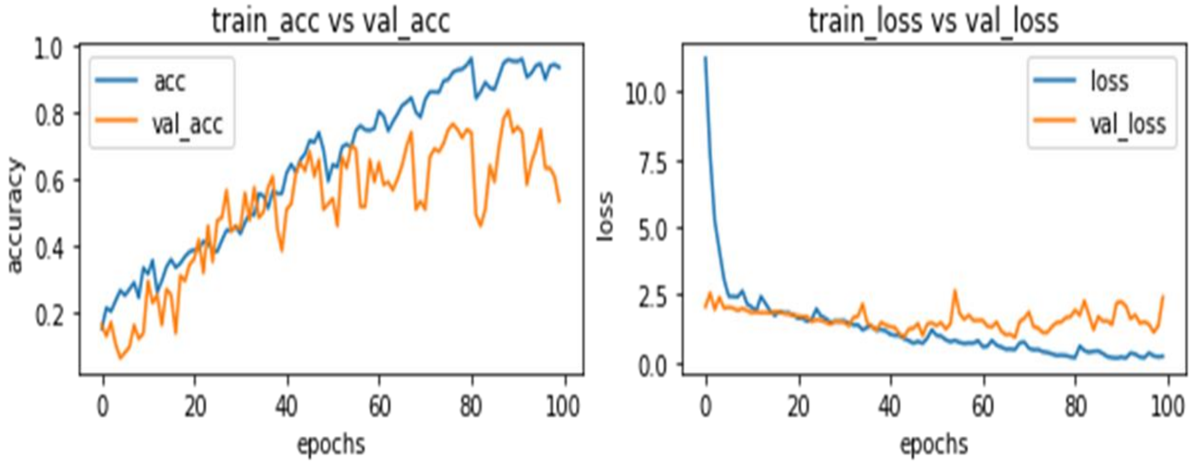


Fig 16: Graph Epoch vs Accuracy and Epochs vs Loss

Accuracy: 0.53

8.3. Our State of art model

Our model is inspired from the research paper written by Octavio Arriaga, Paul G. Plöger, and Matias Valdenegro.

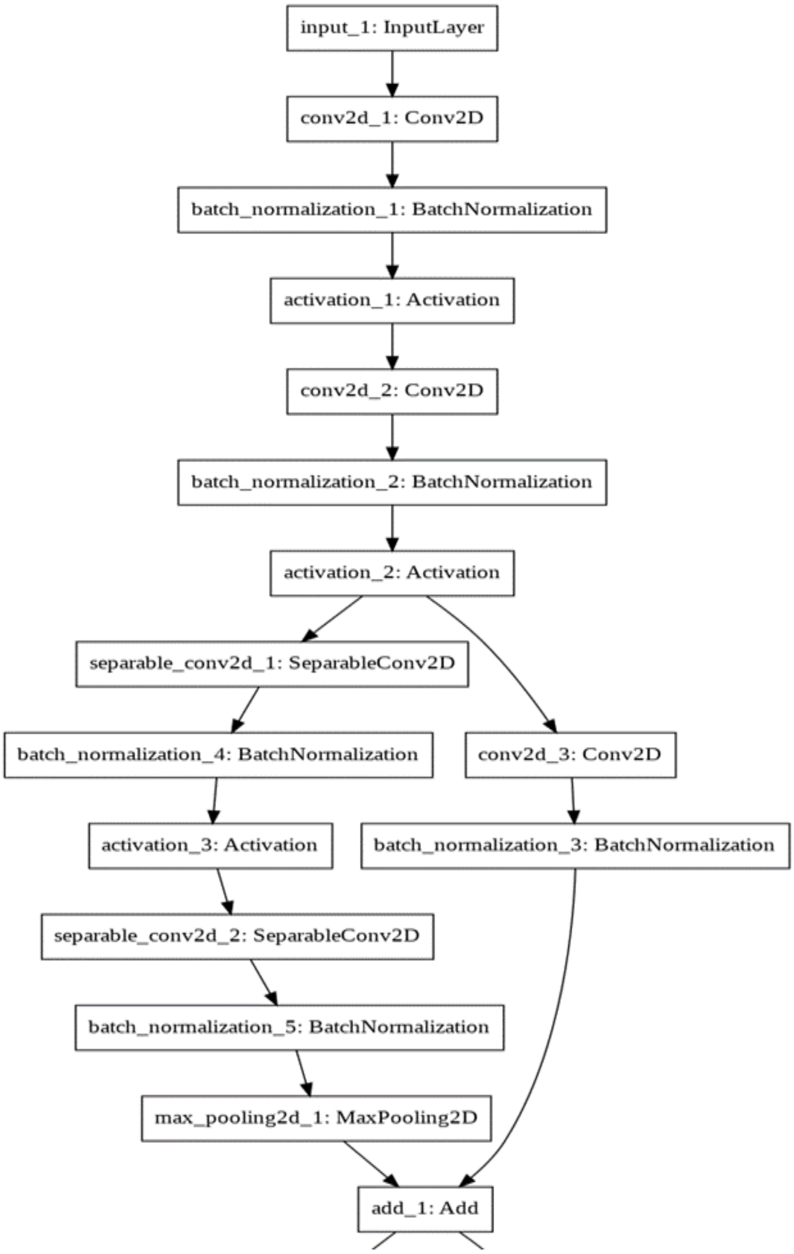

input_1: InputLayer
conv2d_1: Conv2D
batch_normalization_1: BatchNormalization
activation_1: Activation
conv2d_2: Conv2D
batch_normalization_2: BatchNormalization
activation_2: Activation
separable_conv2d_1: SeparableConv2D
batch_normalization_4: BatchNormalization
conv2d_3: Conv2D
activation_3: Activation
batch_normalization_3: BatchNormalization
separable_conv2d_2: SeparableConv2D
batch_normalization_5: BatchNormalization
max_pooling2d_1: MaxPooling2D
add_1: Add

Fig 17: CNN Architecture Block

OUR PROCESS

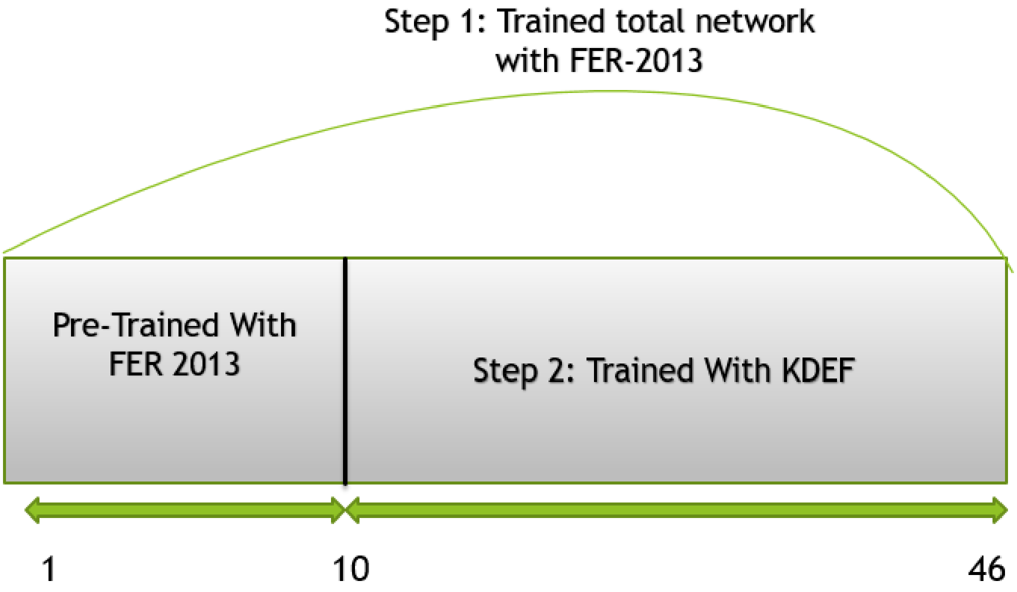
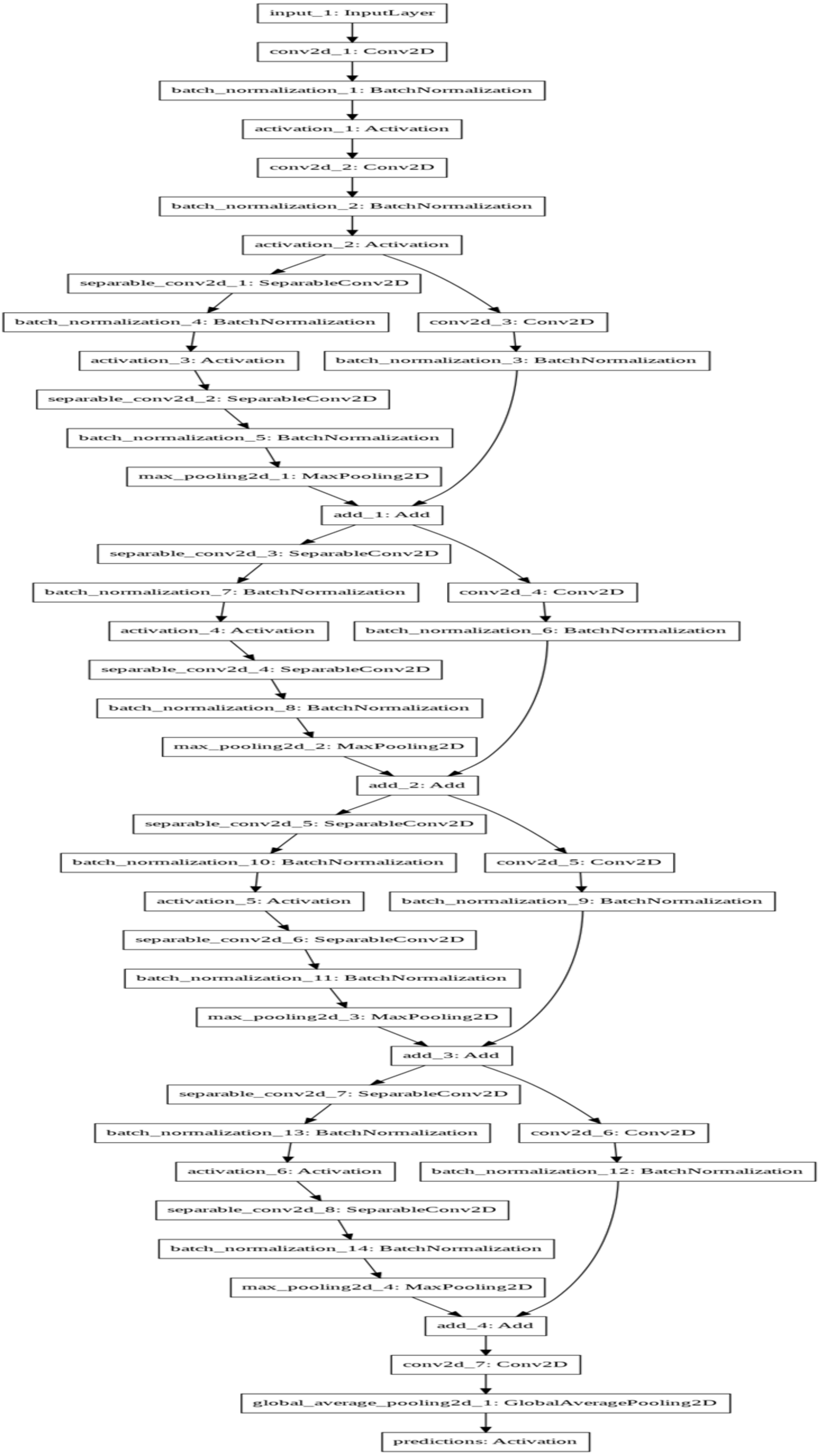
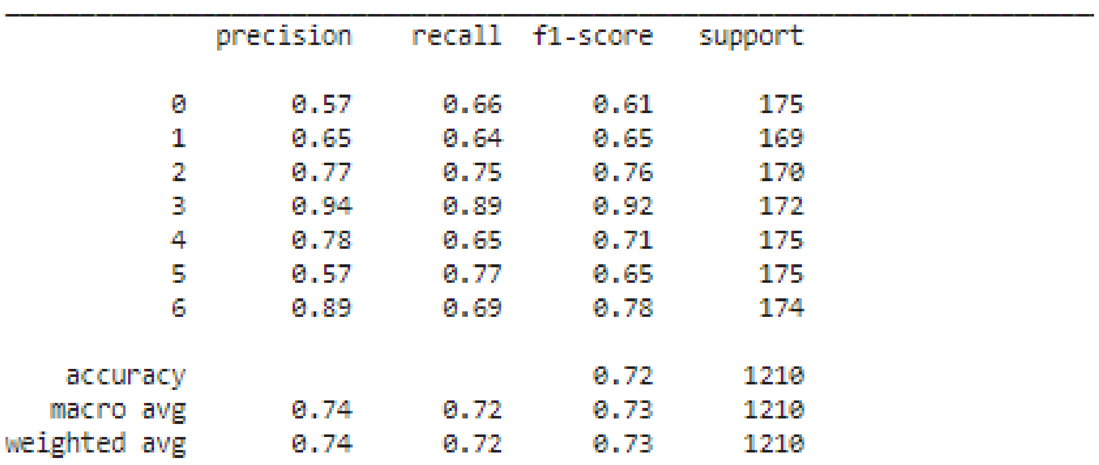

Step 1: Trained total network with FER-2013
Pre-Trained With FER 2013
Step 2: Trained With KDEF
1
10
46

Fig 18: Proposed model structure

Step 1: Trained total network with FER-2013

Step 2: Trained With KDEF


input_1: InputLayer
conv2d_1: Conv2D
batch_normalization_1: BatchNormalization
activation_1: Activation
conv2d_2: Conv2D
batch_normalization_2: BatchNormalization
activation_2: Activation
separable_conv2d_1: SeparableConv2D
batch_normalization_4: BatchNormalization
conv2d_3: Conv2D
activation_3: Activation
batch_normalization_3: BatchNormalization
separable_conv2d_2: SeparableConv2D
batch_normalization_5: BatchNormalization
max_pooling2d_1: MaxPooling2D
add_1: Add
separable_conv2d_3: SeparableConv2D
batch_normalization_7: BatchNormalization
conv2d_4: Conv2D
activation_4: Activation
batch_normalization_6: BatchNormalization
separable_conv2d_4: SeparableConv2D
batch_normalization_8: BatchNormalization
max_pooling2d_2: MaxPooling2D
add_2: Add
separable_conv2d_5: SeparableConv2D
batch_normalization_10: BatchNormalization
conv2d_5: Conv2D
activation_5: Activation
batch_normalization_9: BatchNormalization
separable_conv2d_6: SeparableConv2D
batch_normalization_11: BatchNormalization
max_pooling2d_3: MaxPooling2D
add_3: Add
separable_conv2d_7: SeparableConv2D
batch_normalization_13: BatchNormalization
conv2d_6: Conv2D
activation_6: Activation
batch_normalization_12: BatchNormalization
separable_conv2d_8: SeparableConv2D
batch_normalization_14: BatchNormalization
max_pooling2d_4: MaxPooling2D
add_4: Add
conv2d_7: Conv2D
global_average_pooling2d_1: GlobalAveragePooling2D
predictions: Activation



```
______________________________________________________________________
              precision    recall  f1-score   support

           0       0.57      0.66      0.61       175
           1       0.65      0.64      0.65       169
           2       0.77      0.75      0.76       170
           3       0.94      0.89      0.92       172
           4       0.78      0.65      0.71       175
           5       0.57      0.77      0.65       175
           6       0.89      0.69      0.78       174

    accuracy                           0.72      1210
   macro avg       0.74      0.72      0.73      1210
weighted avg       0.74      0.72      0.73      1210
```

Fig 20: Classification Report

Parameter Used

Loss: Categorical Cross entropy
Optimizer: Adam
Learning Rate: 0.01
Batch size: 64
Epochs: 400

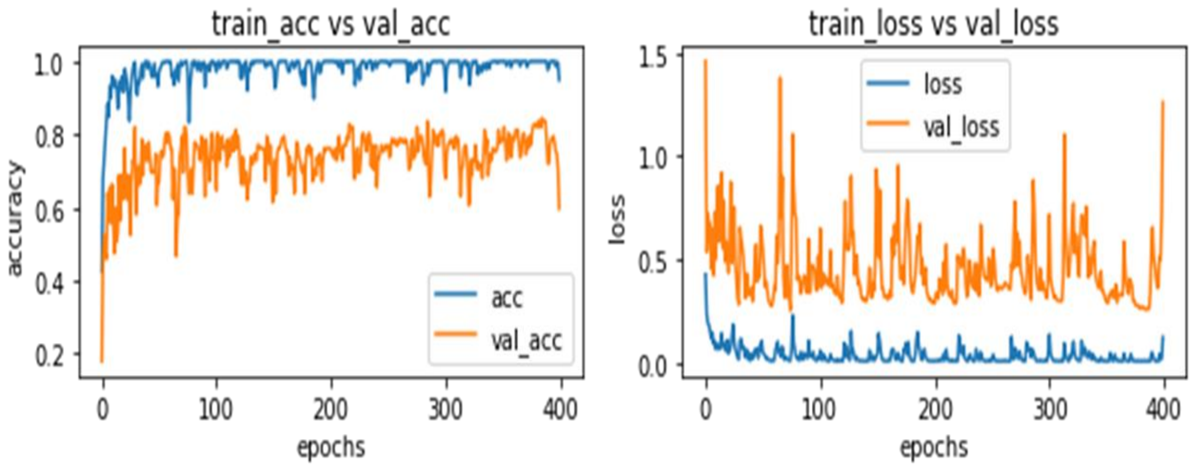


Fig 21: Graph Epoch vs Accuracy and Epochs vs Loss

9. Future work

Future work One of the most obvious continuations of this work is the creation of a multimodal model that combines the video and audio inputs, and therefore significantly improves the results. The next step would be to save the parameters of the model to be able to use it in an application in real-time. For example, this model could be added to a humanoid robot for being able to detect the feelings of the people with whom the robot interacts and improve the performance as a service robot, getting to empathize with humans and trying to make them feel better.

Surely if we applied the models created in a real application, its effectiveness would be deficient, since the models have been created with acted databases and the categories of emotions are very rigid. Probably for a future work the model needs to be tested against different types of data, changing the background and light conditions too. Also, it could be investigated if applying another emotion classification, such as dimensional, which takes into account the intensity of the emotions, could be more effective for recognizing the most intense one and not erring in detecting the most subtle ones

References

1. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput
2. Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE winter conference on applications of computer vision (WACV); 2016. p. 1–10.
3. Zavaschi TH, Britto AS Jr, Oliveira LE, Koerich AL (2013) Fusion of feature sets and classifiers for facial expression recognition. Expert Syst Appl 40(2):646–655
4. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." Journal of personality and social psychology, vol. 17, no. 2, pp. 124–129, 1971.
5. P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique," Psychological bulletin, vol. 115, no. 2, pp. 268–287, 1994.
6. D. Matsumoto, "More evidence for the universality of a contempt expression," Motivation and Emotion, vol. 16, no. 4, pp. 363–368, 1992.
7. R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," Proceedings of the National Academy of Sciences, vol. 109, no. 19, pp. 7241–7244, 2012.
8. Ma, L., Khorasani, K., "Facial expression recognition using constructive feedforward neural networks," Systems, Man, and Cybernetics, Part B: Cybernetics.
9. Liu, P., Han, S., Meng, Z., and Tong, Y., "Facial expression recognition via a boosted deep belief network," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1805–1812.
10. [Online]. Available: https://medium.com/towards-data-science/from-scikit-learn-totensorflow-part-1-9ee0b96d4c85, (last retrieved on: 11/1/2017).
11. [Online]. Available: https://medium.com/towards-data-science/from-scikit-learn-totensorflow-part-1-9ee0b96d4c85, (last retrieved on: 11/1/2017).
12. Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J., "Coding facial expressions with gabor wavelets," in Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. Proceedings, 1998, pp. 200

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

14. Konwar, S. and Borah, S., "ANN based human facial expression recognition in color images," in Proc. International Conference on High Performance Computing and Applications (ICHPCA), 2014.

15. Lopes, A. T., Aguiar, E., and. Santos, T.O., "A facial expression recognition system using convolutional networks," in Proc. 28th SIBGRAPI Conference on Graphics,Patterns and Images, 2015.

16. [Online]. Available: http://www.numpy.org/, (last retrieved on: 11/1/2017).

17. Chollet, F, "Xception: deep learning with separable convolutions." arXiv Prepr. arXiv1610 2357 (2016).

18. Arriaga, Octavio, Matias Valdenegro-Toro, and Paul Plger, "Real-time Convolutional Neural Networks for Emotion and Gender Classification," arXiv preprint arXiv:1710.07557 (2017).

19. Jung H, Lee S, Yim J, Park S, Kim J, "Joint fine-tuning in deep neural networks for facial expression recognition." Proc. IEEE International Conf. on Computer Vision. 2015.

20. https://github.com/Microsoft/FERPlus/.

21. E. Barsoum, C. Zhang, C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution. Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016.

22. K. Simonyan, A. Zisserman, Very deep convolutional networks for largescale image recognition. arXiv preprint arXiv:1409.1556, 2014.

23. Shin, M., Kim, M., and Kwon, D.S., "Baseline CNN structure analysis for facial expression recognition," in Proc 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2016.

24. Goeleven E., DeRaedt R., Leyman L., and Verschuere B, The Karolinska directed emotional faces: a validation study, Cognition and Emotion, 22(6), 1094-1118,2008.