Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 11. No. 4, December, 2023

Article Citation Format
Oyedeji, Ayo I., Adenle, Bamidele J. & Akinrogunde, Oludare O.. (2023): Facial Emotion Detection: A Comprehensive Exploration of Convolutional Neural Networks. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 11, No. 3. Pp 1-30
dx.doi.org/10.22624/AIMS/DIGITAL/V11N4P1
www.isteams.net/digitaljournal.

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 17th September, 2023
Review Type: Blind Peer
Final Acceptance: 12th October, 2023

Facial Emotion Detection: A Comprehensive Exploration of Convolutional Neural Networks

[1]Oyedeji, Ayo I., [2]Adenle, Bamidele J. & [3]Akinrogunde, Oludare O.
[1]Department of Computer Engineering, Ogun State Institute of Technology, Igbesa
[2]Department of Software Engineering, Dots ICT Institute of Technology, Abeokuta, Ogun State
[3] Department of Electrical and Electronics Engineering, Institute of Technology, Igbesa ,Ogun
E-mails: ayooyee@ogitech.edu.ng, bamidelejohnson2@gmail.com, akinrogunde2015@gmail.com

ABSTRACT

Facial emotions play a crucial role in non-verbal communication, as they reflect the internal feelings of individuals through expressions on their faces. Recognizing and interpreting these facial expressions have significant applications in various fields, especially human-computer interaction. In this journal, a facial emotion detection system based on convolutional neural networks (CNN) was developed. The primary objective was to classify facial images into different emotional categories. The CNN models were trained using grayscale images, and the training process was optimized by leveraging GPU computation. To accommodate new subjects efficiently, the last two layers of the CNN were selectively trained, reducing the overall training time. Image preprocessing steps were implemented in MATLAB, while the CNN algorithm was implemented in C language using the GCC compiler. A user-friendly Graphical User Interface (GUI) in MATLAB seamlessly integrated all the processing steps, from image preprocessing to facial emotion detection. The performance evaluation was conducted using the FER2013 dataset, achieving an impressive accuracy of 78.2% with an average training time of less than 10 minutes when incorporating 1 to 10 new subjects into the system. This work demonstrates the effectiveness of CNN-based approaches for accurate and efficient facial emotion detection, offering promising results in real-world applications.

Keywords: Neural Network, Convolutional, GUI, Emotion, Facial, Algorithm

1. INTRODUCTION

Facial expressions play a key role in understanding and recognizing emotions. Even the term “interface” suggests the importance of the face in communication between two entities. Studies have shown that reading facial expressions can dramatically alter the interpretation of what is being said and control the flow of Conversation.[4]

A person's ability to interpret their emotions is very important for Effective communication. The proportion of up to 93% of communication used in a normal conversation depends on the emotion of an entity. For ideal Human Machine Interfaces (HCI), it would be desirable for machines to be able to read human emotions. This research focuses on how systems can correctly detect the emotions of their various sensors. This experience was used as a face image as a means of reading human emotions. Research on human emotions dates back to Darwin's pioneering work and has since attracted many researchers to the field. Seven basic emotions are universal for humans. Namely, neutral, angry, disgusted, fearful, happy, sad and surprised, and these basic emotions can be identified from a person's facial expression. This study suggests an effective way to identify these emotions.

Facial emotion detection system is being identified as an effective way of identifying or perceiving a person's reactions towards an event. Various methods of recognizing emotions have been proposed in recent decades. Many algorithms have been proposed to develop system applications capable of very well detecting facial emotions. Computer applications could communicate better by altering reactions in various interactions depending on the emotional state of human user. A person emotion can be determined by the tongue, face and even gesture. The work presented in this paper explores the recognition of expressions from face. This is all actionable information that organization and businesses can use to understand their customer and create products that people like.[10]. Neural network is a robust machine learning method in pattern recognition field. Neural network has the ability to deal with non-linear problems of data samples. Typical neural network technique is multilayer perceptron (MLP). However, designing a pattern recognition system using this type of neural network ends up with massive interconnection nodes that produce a totally flat structure with inputs are fully connected to the subsequent layers in the architecture. In addition, the topology of the input data is completely ignored, yielding similar training results for all permutations of the input vector.

Another variant of MLP is called Convolutional Neural Network (CNN) that is proposed by Yann Lecun in 2014 through LeNet-5 architecture. CNN has been applied to wide range of applications including face detection and recognition, gender recognition, object recognition, etc. CNN method has been used as a face recognition technique through several existing works. Most of the works produce complex CNN design that incurs in additional cost and increasing the training time.[4] Convolutional neural networks are deep artificial neural networks. We use CNN to classify images, cluster them by similarity (photo search), and perform object recognition within scenes. It can be used to identify faces, individual, street signs, tumors, platypuses and many other aspects of visual data. The convolutional layer is the core building block of a CNN.

The layer's parameters consist of a set of learnable filters (or kernels) which have a small receptive field but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product, and producing a 2-dimensional activation map of that filter. As a result, the network learns about the filters. The filter activates when they see some specific type of feature at some spatial position in the input. Then the activation maps are fed into a down sampling layer, and like convolutions, this method is applied one patch at a time. CNN has also fully connected layer that classifies output with one label per node.[10]

The work presented in this project examines the detection of facial emotion. For facial emotion detection, with the convolutional neural network being used, the system is designed using a software application called MATLAB on which the system will be built and trained. MATLAB is an object-oriented high level interactive software package for scientific and engineering numerical computations for analyzing and designing of systems and products that transform our world. It enables easy manipulation of matrix and other computations without the need for traditional programming. According to (Roman Randil, 2017) said Convolutional Neural Networks (CNN) are very similar to ordinary Neural Networks. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other.[7]

There is controversy surrounding the question of whether facial expressions are a worldwide and universal display among humans. Supporters of the Universality Hypothesis claim that many facial expressions are innate and have roots in evolutionary ancestors. Opponents of this view question the accuracy of the studies used to test this claim and instead believe that facial expressions are conditioned and that people view and understand facial expressions in large part from the social situations around them. Moreover, facial expressions have a strong connection with personal psychology. Some psychologists have the ability to discern hidden meaning from a person's facial expression. In Fard et al. (2022) the network was guided by an Ad-Corre Loss which showcase the intact feature and their corresponding vectors with strong correlatedness for inter-class samples.

The decreased correlation for the in- between-class samples are also checked for correlations. There exist three components of Ad-Corre. Vulpe- Grigoraşi et al. (2021) proposed a strategy for improving the accuracy of face emotion identification was implemented by tweaking the hyperparameters of a convolutional neural network. The network's ideal hyperparameters were obtained by creating and training models using the Random Search method on a search space specified by discrete hyperparameter values. A CNN model for image categorization focusing on principal component analysis initialization was suggested by Ren et al.(2016).Rather than employing PCA to reduce down the dimensionality of the initial input, then PCA is made to work in effectiveness with the CNN training procedure. Such initialization values can lessen the influence of gradient diffusion problems caused by incorrect starting settings by containing image-altering information. A 2020 study on "emotion residue" found that even when study participants attempted to make neutral facial expressions, their faces still retained emotion residue from prior expressions, and these prior expressions were able to be detected by observers.

They pre-trained of models on the FER-2020 dataset and then finetuned the model on the Static Facial Expressions in the Wild 2.0 (SFEW) dataset. They used an ensemble of three face detectors to detect and extract faces from the labelled movie frames of SFEW. They then proposed a data perturbation and voting method to increase the recognition performance of the CNN. They also chose to use stochastic pooling layers over max pooling layers citing its better performance on their limited data. Kahou et al. used a CNN-RNN architecture to train a model on individual frames of videos as well as static images They made use of the Acted Facial Expressions in the Wild (AFEW) 5.0 dataset for the video clips and a combination of the FER-2013 and Toronto Face Database for the images Instead of using long short-term memory (LSTM) units, they used IRNNs which are composed of rectified linear units (ReLUs). These IRNNs provided a simple mechanism for dealing with the vanishing and exploding gradient problem

They achieved an overall accuracy of 0.528. Mollahosseini et al.Proposed a network consisting of two convolutional layers each followed by max pooling and then four Inception layers. They used this network on seven different datasets including the FER-2013 dataset. They also compared the accuracies of their proposed network with an AlexNet network trained on the same datasets. They found that their architecture had better performance on the MMI and FER-2020 datasets with comparable performances on the remaining five datasets. The FER-2020 dataset in particular managed to reach an accuracy of 0.764. Ming Li et al. propose a neural network model to overcome two shortcomings in still image based FERs which are the inter-variability of emotions between subjects and misclassification of emotions. The model consists of two convolutional neural networks - the first is trained with facial expression databases whereas the second is a Deep ID network used for learning identity features.

These two networks are then concatenated together as a Tandem Facial Expression of TFE Feature which is fed to the fully connected layers to form a new model. The proposed model was evaluated on two datasets, namely the FER+ database and the Extended Cohn Kanade (CK+) database. The identity features were learned from the CASIA-WebFace database. The model was trained for 200 epochs and achieved an accuracy measure of 71.1% on the FER2013 dataset, 99.31% on the CK+ database. These experimental results show that the model outperforms many state-of-the-art methods on the CK+ and FER+ databases.

Tan et al. Propose a neural network model to classify a group image into a particular emotion - positive, neutral or negative. The model consists of two convolutional neural. networks - the first is based on group images and the second is based on individual facial emotions. The facial emotion CNN comprises of two CNNs - one for aligned faces which trained using the ResNet64 model using the Web face dataset and the other for non-aligned faces which is trained the ResNet34 model on the FER+ dataset. The group images are trained using VGG19 model on the Places and ImageNet datasets. Fine-tuning is done with batch normalization average pooling of the ResNet101 and BN-Inception models with a dropout of 0.5. The validation set consists of images - combined from all of the datasets used for training and the model achieved an accuracy measure of 80.9.

1.1 Background to the Study

A facial emotion is one or more motions or positions of the muscle beneath the skin of the face. According to one set of controversial theories, these movements convey the emotional state of an individual to observers. Facial expressions are a form of nonverbal communication. They are a primary means of conveying social information between humans but they also occur in most other mammals and some other animal species Humans can adopt a facial expression voluntarily or involuntarily, and the neural mechanisms responsible for controlling the expression differ in each case. Rasha Talib. (2023) said Voluntary facial expressions are often socially conditioned and follow a cortical route in the brain. Conversely, involuntary facial expressions are believed to be innate and follow a subcortical route in the brain.

Facial recognition is often an emotional experience for the brain and the amygdala is highly involved in the recognition process. The eyes are often viewed as important features of facial expressions. Aspects such as blinking rate can possibly be used to indicate whether a person is nervous or whether he or she is lying. Also, eye contact is considered an important aspect of interpersonal communication. However, there are cultural differences regarding the social propriety of maintaining eye contact or not.

Beyond the accessory nature of facial expressions in spoken communication between people, they play a significant role in communication with sign language Many phrases in sign language include facial expressions in the display. When facial emotion detection system is considered; two main different approaches, both of which include two different methodologies, exist. Dividing the face into separate action units or keeping it as a whole for further processing appears to be the first and the primary distinction between the main approaches. In both of these approaches, two different methodologies, namely the 'Geometric-based' and the 'Appearance-based' parameterizations, can be used. In the following subtitles, details of the two approaches and the two methodologies have been further discussed.

The two main approaches can be summarized as follows: Making use of the whole frontal face image and processing it in order to end up with the classifications of 7 universal facial expression prototypes: disgust, fear, happy, surprise, sadness, neutral and anger;

First Approach

The approach starts by utilizing CNNs, a powerful deep learning architecture known for its ability to learn hierarchical representations from raw input data. The authors design a CNN model specifically tailored for facial expression recognition. The network architecture is carefully crafted to capture the intricate features and patterns associated with different facial expressions. After training the CNN model, the authors evaluate its performance using various metrics, such as accuracy, precision, recall, and F1 score. The model's ability to correctly classify facial expressions is assessed by comparing its predictions with the ground truth labels in a separate testing dataset. The results demonstrate the effectiveness of the proposed approach in accurately recognizing facial expressions. The CNN model achieves high recognition accuracy, showcasing its potential for practical applications in domains such as human-computer interaction, emotion analysis, and facial expression-based systems

Second Approach

This approach has been presented to be the 'Facial Action Coding System', which is first developed by Ekman and Friesen (Ekmen & Friesen, The Facial Actin Coding System: A Technique for the Measurement of Facial Movement, 2017), for describing facial expressions by 44 different Action Units (AU's) existing on face. The advantage is that; this decomposition widens the range of applications of face expression recognition. This is due to ending up with individual features to be used in/with different processing areas/method so the than just having the 6 universal facial expression prototypes. Most of the current work done on facial expression analysis makes use of these action units. It should be mentioned that, there are also some other methods in which neither the frontal face image as a whole nor the all of 44 action units themselves, but some other criterion such as the manually selected regions on face (Mase, 2009) or surface regions of facial features (Yacoob & Davis, 2016) are used for the recognition of the facial expression. There are two main methods that are used in both of the above explained approaches:

Geometric Based Parameterization: is an old way which consists of tracking and processing the motions of some spots on image sequences, firstly presented by Suwa to recognize facial expressions (Suwa, Sugie, & Fujimora, 2017). Cohn and Kanade later on tried geometrical modeling and tracking of facial features by claiming that each AU is presented with a specific set of facial muscles. In general, facial motion parameters (Mase, 2019) (Yacoob & Davis, 2018 and the tracked spatial positioning & shapes of some special points (Lanitis, Taylor, & Cootes, 2007) (Kapoor, Qi, &

Picard, 2003) on face, are used as feature vectors for the geometric based method. These feature vectors are then used for classification. The following might be regarded as the disadvantages of this method: The approximate locations of individual face features are detected automatically in the initial frame; but, in order to carry out template-based tracking, the contours of these features and components have to be adjusted manually in this frame. (And this process should be carried out for each individual subject). The problems of robustness and difficulties come out in cases of pose and illumination changes while the tracking is applied on images.

As actions & expressions tend to change both in morphological and in dynamical senses, it becomes hard to estimate general parameters for movement and displacement. Therefore, ending up with robust decisions for facial actions under these varying conditions become to be difficult (Donato, Bartlett, Ekmen, & Sejnowksi, 2009). Rather than tracking spatial points and using positioning and movement parameters that vary within time, color (pixel) information of related regions of face are processed in **Convolutional Neural Network (CNN)**

A Convolutional neural network is a neural network comprised of convolution layers which does computational heavy lifting by performing convolution. Convolution is a mathematical operation on two functions to produce a third function. It is to be noted that the image is not represented as pixels, but as numbers representing the pixel value. In terms of what the computer sees, there will simply just be a matrix of numbers. The convolution operation takes place on these numbers. We utilize both fully-connected layers as well as convolutional layers. In a fully-connected layer, every node is connected to every other neuron.

They are the layers used in standard feed forward neural networks. Unlike the fully connected layers, convolutional layers are not connected to every neuron. Connections are made across localized regions. A sliding" window" is moved across the image. The size of this window is known as the kernel or the filter. They help recognize patterns in the data. For each filter, there are two main properties to consider - padding and stride. Stride represents the step of the convolution operation, that is, the number of pixels the window moves across. Padding is the addition of null pixels to increase the size of an image. Null pixels here refer to pixels with value of 0. If we have a 5x5 image and a window with a 3x3 filter, a stride of 1 and no padding, the output of the convolutional layer will be a 3x3 image. This condensation of a feature map is known as pooling. In this case," max pooling" is utilized. Here, the maximum value is taken from each sliding window and is placed in the output matrix.

Convolution is very effective in image recognition and classification compared to a feed-forward neural network. This is because convolutional neural network reduces the number of parameters in a network and take advantage of spatial locality. Further, convolutional neural networks introduce the concept of pooling to reduce the number of parameters by down-sampling. Applications of Convolutional neural networks include image recognition, self-driving cars and robotics. CNN is popularly used with videos, 2D images, spectrograms, Synthetic Aperture Radars.

Approach includes:

- As a preprocessing step for the CNN, also apply Gaussian filter to the images, and subtract the mean-image of the training set from each image. In order to get more out of our limited training data, also augment the images to include reflections and rotations of each image, with the hope that it would improve robustness
- **Develop real-time Interface**: Open CV allows us to get images from our laptop's webcam. Then extract the face as before, pre-processed the image for the CNN, and sent it to AWS. On the server, a script would run the image through the CNN, get a prediction and the results would be pulled back to local.

One future area of work is to create a user interface where users can iteratively train the model through correcting false labels. This way the model can also learn more from real world users who express various emotions in different ways. In addition, including a layer in the network that account for class imbalance could provide addition improvements over the results.

Performance Metrics In order to measure the performance rate of a biometric system, several parameters or metrics are considered. Some of the performance metrics are: Accuracy, Precision, Epoch, Iteration, Batch-loss, learning rate etc.

2. RELATED WORK

Generally, there are lot efforts that have been done in the field of Facial Emotion Detection system. However, this field is still an active research area and more efforts are being done to improve the accuracy, efficiency and effectiveness of the emotion detection systems. Yu and Zhan used a five-layer ensemble CNN to achieve 0.612 accuracy. They pre-trained their models on the FER-2013 dataset and then fine-tuned the model on the Static Facial Expressions in the Wild 2.0 (SFEW) dataset. They used an ensemble of three face detectors to detect and extract faces from the labeled movie frames of SFEW.

They then proposed a data perturbation and voting method to increase the recognition performance of the CNN. They also chose to use stochastic pooling layers over max pooling layers citing its better performance on their limited data. Kahou et al. used CNN-RNN architecture to train a model on individual frames of videos as well as static images. (2020). They made use of the Acted Facial Expressions in the Wild (AFEW) 5.0 dataset for the video clips and a combination of the FER-2013 and Toronto Face Database for the images. Instead of using long short term memory (LSTM) units, they used IRNNs which are composed of rectified linear units (ReLUs). These IRNNs provided a simple mechanism for dealing with the vanishing and exploding gradient problem. They achieved an overall accuracy of 0.528.

Mollahosseini et al. proposed a network consisting of two convolutional layers each followed by max pooling and then four Inception layers. They used this network on seven different datasets including the FER-2013 dataset. They also compared the accuracies of their proposed network with an AlexNet network trained on the same datasets. They found that their architecture had better performance on the MMI and FER-2013 datasets with comparable performances on the remaining five datasets. The FER-2013 dataset in particular managed to reach an accuracy of 0.664.

Ming Li et al. propose a neural network model to overcome two shortcomings in still image based FERs which are the inter-variability of emotions between subjects and misclassification of emotions. The model consists of two convolutional neural networks - the first is trained with facial expression databases whereas the second is a Deep ID network used for learning identity features. (Manednis 2020) said These two networks are then concatenated together as a Tandem Facial Expression of TFE Feature which is fed to the fully connected layers to form a new model. The proposed model was evaluated on two datasets, namely the FER+ database and the Extended Cohn- Kanade (CK+) database. The identity features were learned from the CASIA-WebFace database. The model was trained for 200 epochs and achieved an accuracy measure of 71.1% on the Wang, Y. (2020). FER2013 dataset, 99.31% on the CK+ database. These experimental results show that the model outperforms many state-of-the-art methods on the CK+ and FER+ databases. Tan et al. propose a neural network model to classify a group image into a particular emotion - positive, neutral or negative. The model consists of two convolutional neural networks.

Most other works in the same field attempted to solve the facial emotion recognition problem by the use of a combination of different datasets. In this paper, a single dataset, FER-2013 was chosen over such a combination of different datasets and then experiments were conducted to find the highest accuracy the model could reach. Although there is much previous recognition system for faces, three categories of analyses would be made in this section; some of the systems that used various methods, some of the systems that used CNNs, and some of the systems
which have been evaluated on FEI dataset.

The researchers in propose face image recognition methods using discrete complex fuzzy transform (DCFT) and local binary pattern (LBP). They also used maximum pooling technology to create the feature fusion method. They used 5 facial datasets to prove the effectiveness of their proposed method and tested using SVM and KNN classifiers. The authors in examined extensive works for face recognition development in the recent era. They described the techniques for face recognition in detail including various datasets. They analyzed the datasets including benchmark

accuracy results and specific information. The various techniques for face detection are also studied and discussed. The availability of recognition methods is described in precisely and debate for those approaches. The authors in proposed a new approach for recognition of the faces using Bayesian Neural Networks recognizers. They extract face information for identification using Gray Level Co-occurrence Matrix. They used the ORL database to the MLP of one hidden layer that has 10 hidden neurons. The input is 6 and the output has 40 neurons. They used the SoftMax function for output units. They get the 81.23 % for the classification test.

The researchers in proposed a technique to detect and recognition of the face using image processing methods. They detect faces using MatLab toolbox and PCA features are used. They assumed about 90% as the classification accuracy would get. The combination features of spatial domain and transformation domain are invented in. They extract the face using popular Viola Jones Algorithm and scaled to 100x100 size. Then spatial feature is extracted using Asymmetric Region Local Binary Pattern (ARLBP) and transformation features are extracted using Fast Discrete Curvelet Transform (FDCT). Then these two features are concatenated to classify the input image against the trained images. They tested for JAFFE, L-Space k, FERET and NIR databases and get a recognition rate between 76.8 % to 95.48 %.

After the recent various face recognition system based on different implementation methods have been analyzed, the different CNNs based methods are started to discuss. The author in worked extensive reviews for deep learning face recognition system. The deep deep neural network based on CNNs has been proposed in.[13] They used the whole pixel data as the input image and using Yale faces dataset to test the system with a 9:1 ratio and achieved 97.05%. They used 4 layers and the results are produced using 50 epochs.

The systematic evolution for face recognition using deep learning is summarized in. [6] They made a comprehensive evolution rather than a normal survey among the DCNN implementation diversities. They used the combination of CASIA-WebFace and UMD-Faces named UMD-CASIA as the 90% training and 10% validation set. They created YTF, CACD-VS, LFW, and CFP datasets are as the testing set.

Table 1.1 facial emotion with various corresponding expression

Emotion	Facial Expression
Anger	Lowered and burrowed eyebrows Intense gaze Raised chin
Happy	Raised corners of mouth into a smile
Surprise	Dropped jaw Raised brows Wide eyes
Fear	Open mouth Wide eyes Furrowed brows
Sadness	Furrowed brows Lip corner depressor
Disgust	Biting of the lips

They used MTCNN as the face detector and analyzed for all criteria of CNN including batch normalization, feature normalization, down sample, and SE block. They tested on ResNet-50, Face-ResNet, Google-Net, and VGG-16 networks for all datasets and evaluate the results. They showed that ResNet-50 provides the best results among the DCNN.

The system for face recognition base on CNN is proposed in where they applied the Dropout idea to scale the activation values. They use 20,000 face images of private university datasets called SWUNs which are taken from 50 peoples. They normalize the face image into 32x32 gray images to input deep CNN and get 98.8% for recognition. They showed that the scaling of the activation values for training and testing can improve the classification results.

Fig 1. Facial Expressions for Different Emotions

3. FINDINGS AND DISCUSSION

In this project work, a facial image captured by digital camera was passed into the system for classification. Noise and other unwanted elements are removed from the image, that is, conversion of face images into grayscale and application of local histogram equalization for enhancement contrast as illumination normalization.[14] Convolutional neural network algorithm was applied in the training of the system. Finally, classification of individual images based on input image was tested using CNN classifier. The component diagram is shown below in fig. 2

Acquisition of Face Images

Face images with 28,821 facial images consisting of 3993 angry face, 436 disgust face, 4103 fear faces, 7164 happy faces, 4982 neutral faces, 4938 sad faces and 3205 surprised faces expressions gotten from an online dataset source (FER2013) named kaggle.com in the size 48 by 48 pixels. The original face images will be downsized without any alteration in the images. All face images taken have equal uniform illumination conditions and light color background.

Image pre-processing

In this phase, image pre-processing was carried out by converting face images into grayscale, application of illumination normalization such as local histogram equalization for enhancement contrast and by calculating the average face vector and subtracting average face from each face vector. [9] This removes noise and other unwanted element from the face images.

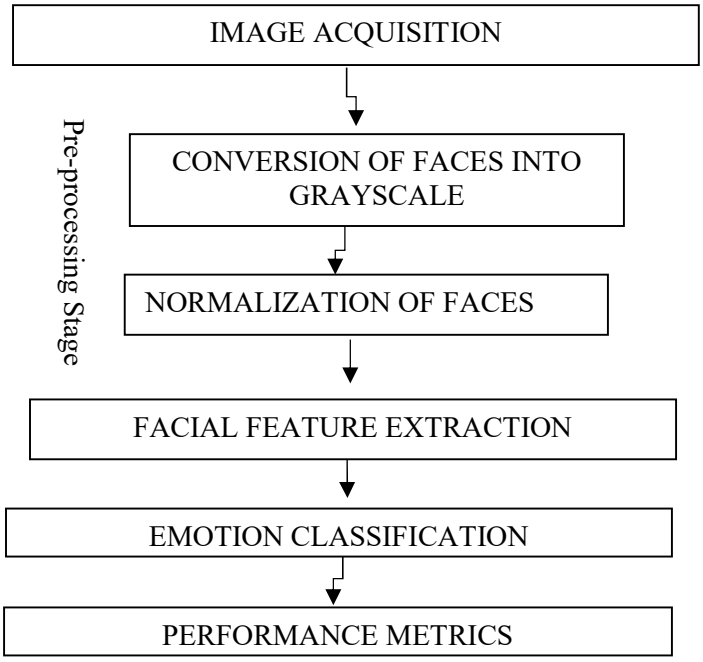

Figure 2. Components of a Facial Emotion Detection System

Conversion of Face Images into grayscale and Face Vector

The images acquired from the digital camera are color images (R, G, B) and were converted into grayscale with pixel value between 0 and 255, that is, image in black and white. Each of the grayscale images were expressed and stored in form of matrix in MATLAB which eventually will be converted to Vector image for further processes. The lightness method of converting color to grayscale was employed and averages the most prominent and least prominent colors. It will be denoted by X as shown in equation 3.1.

$$X = \left(\frac{\max(R,G,B)+\min(R,G,B)}{2}\right) \qquad (3.1)$$

Normalization of Face Image

Normalization removes any common features that all face images shared together, so that each face images can be left with unique features. The common feature was discovered by finding the average face vector of the whole training set (face images). [8] Then, the average face vector will be subtracted from each of the face vectors which resulted to normalized face vector. Histogram equalization will be used for enhancement contrast that ensures that the input pixel intensity, x is transformed to new intensity value (x') by T as shown in the equation below. The transform function (T) is the product of a cumulative histogram and a scale factor. The scale factor is needed to fit the new intensity value within the range of the intensity values.

$$x' = T(x) = \sum_{i=0}^{r} n_i \cdot \frac{\max intensity}{N} \qquad (3.2)3$$

where n_i is the number of pixels at intensity i, N is the total number of pixels in the image.

Feature Extraction

Principal Component Analysis was used as feature extraction which converted the set of correlated face images into set of uncorrelated eigenfaces and was also used for dimension reduction of the face vector space. Dimensional reduction is the transformation of normalized face vector space into lower dimensional subspace, that is, the dimensionality of the original training set is reduced before eigenfaces are calculated. Eigenfaces (eigenvectors) are the principal components of the training set of face images generated after reducing the dimensionality of the training set. PCA eigenface method considers each pixel in an image as a separate dimension, that is, N by N image has N^2 pixels or N^2 dimensions. To calculate eigenvector, there is a need to calculate the covariance matric C as shown in Equation 3.3.

$$C = A \,.\, A^T \tag{3.3}$$

where,

$$A = N^2 \; by \; M \tag{3.4}$$

where N in Equation 3.4 is the dimension of the image, M is the number of column vector. If eigenvector will be calculated from a covariance matrix before dimension reduction, the system would slow down terribly or run the system out of memory, due to huge computations. In order to overcome this problem, the solution is to calculate eigenvectors from the covariance matrix with reduced dimensionality. Therefore, the covariance matrix was calculated for the eigenvector in the inverse form as in equation 3.5.

$$C = A^T \,.\, A \tag{3.5}$$

where,

$$A = M \; by \; N^2 \tag{3.6}$$

This gives room for dimension reduction. The eigenvectors will be sorted according to their corresponding eigen values from high to low. Then, the eigenvectors corresponding to zero eigen values are discarded while those associated with non-zero eigen values are kept. Consequently, the Eigen face is formed.

Classification Using Convolutional Neural Network

Convolutional Neural Network is not just a deep neural network that has many hidden layers. It is a deep network that imitates how the visual cortex of the brain processes and recognizes images.
The input image enters into the feature extraction network. The extracted feature signals enter the classification neural network. The classification neural network then operates based on the features of the image and generates the output. The feature extraction neural network consists of piles of the convolutional layer and pooling layer pairs. The convolution layer, as its name implies, converts the image using the convolution operation. It can be thought of as a collection of digital filters. The pooling layer combines the neighboring pixels into a single pixel. Therefore, the pooling layer reduces the dimension of the image. As the primary concern of Convolutional Neural Network is the image; the operations of the convolution and pooling layers are conceptually in a two-dimensional plane. This is one of the differences between Convolutional Neural Network and other neural networks.

The framework design of our emotion detection by facial expressions contains training and real-time processes. The training process's input is a facial expression dataset that was collected from actors who practiced facial expressions with different emotions to naturally express emotions. Then, the face region of each input is detected to extract facial features to form training feature vectors for the classification process, which learns the training input based on selected classification methods to constructs a classifier to recognize emotions in real-time processing. The workflow of the real-time emotion detection by facial expressions proceeds as follows.

- The emotion detection finds a user's face from the video frames (input).
- The detection extracts the facial features and normalizes them to form feature vectors.
- It then classifies the user emotions into one of seven classes (neutral, happiness, sadness, anger, disgust, fear and surprise) using a classifier that is generated from training process.
- Finally, it calculates the percentage of each emotion for further analysis.

Implementation of Facial Emotion Detection System in MATLAB

An interactive Graphic User Interface (GUI) was developed with a real time database consisting of more than 28000 subjects of face images. The implementation tool used was MATLAB R2018, a version on Windows 10 professional 64-bit operating system, Intel®Corei5® CPU B960@2.70GHZ Central Processing Unit, 8GB Random Access Memory and 500GB hard disk drive.

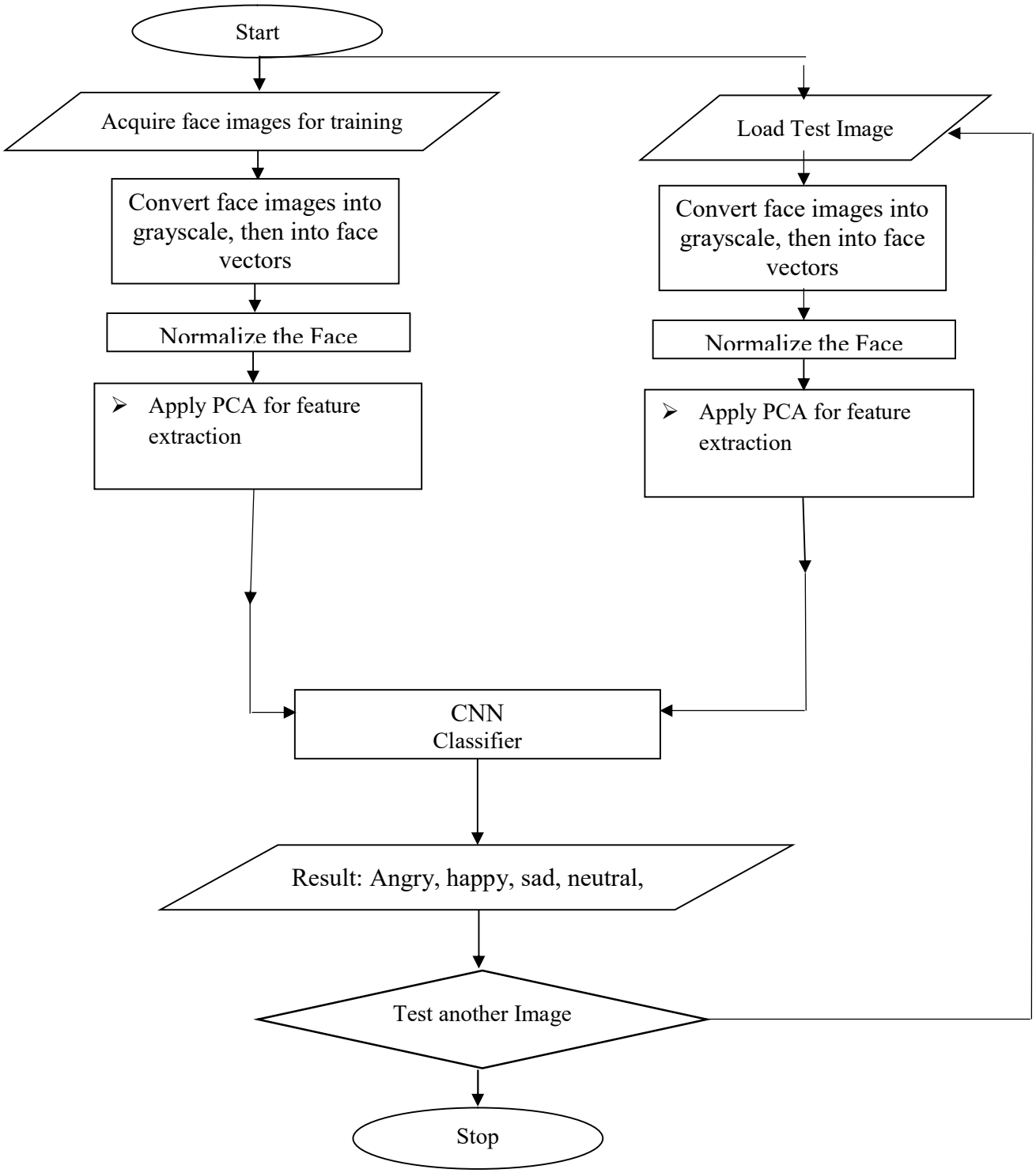

Figure 3: Flowchart Showing Trained and Tested Faces

Face Emotion Detection Stages/Phases

The facial recognition process has five phases or steps. The phases are Image Acquisition, Image Processing, Feature Extraction Technique, Feature Selection Technique and Emotion Classification Technique. These steps are separate components of a facial recognition system and depend on each other (Marques, 2010), (Lucas and Helen, 2009). Figure 2.1 present the relationship diagram between the phases.

Performance Metrics

The performance of Convolutional Neural Network (CNN) on trained and detected emotion was measured using epoch, Iteration, Time Elapsed, Mini-batch accuracy, Base Learning rate.

The following parameters are used to measure or evaluate the overall performance of the system:

- **Epoch:** An **epoch** is a term used in machine learning to indicate the number of passes of the entire training dataset the machine learning algorithm has completed. Datasets are usually grouped into batches (especially when the amount of data is very large).
- **Iteration rate: Iteration** is a term used in machine learning and indicates the number of times the algorithm's parameters are updated. Training of a neural network will require much iteration.
- **Time Elapsed**: Elapsed time is the amount of time that passes from the start of an event to its finish. In simplest terms, elapsed time is how much time goes by from one time
- **Mini-batch accuracy:** The mini-batch accuracy reported during training corresponds to the accuracy of the particular mini-batch at the given iteration. It is not a running average over iterations.

Merit of Convolutional Neural Network

- It automatically detects important features without human supervision
- It has numerical strength
- It has the ability to work with inadequate knowledge
- It has fault tolerance
- It has ability to train machine

Demerit of Convolutional Neural Network

- It is significantly slow due to operations such as maxpool
- Unexplained functioning of the system. It is expensive

4. RESULT

The dataset contains 28821 Face images consisting of 3993 angry face, 436 disgust face, 4103 fear faces, 7164 happy faces, 4982 neutral faces, 4938 sad faces and 3205 surprised faces expressions, which was used in training and testing of the system model. All images were used for training and testing using Convolutional Neural Network algorithm implemented on MATLAB software. The images dataset was divided into 2 datasets; training and testing (80% of the dataset was used for the training of the system and the remaining 20% was used for the testing of the system).

The list of main faces is then down-sampled to generate a final list containing 5764 faces for testing. Features are then extracted from each face in the list using a pre-trained VGG16 model based on the VGGFace dataset. The face features from each face are concatenated to create a single feature vector. Finally, a Convolutional Neural Network (CNN) classifier is trained to classify image emotions using the Adam optimizer for higher accuracy. During training, the weights of the pre-trained models were frozen, implying that only the weights of the CNN were adjusted. We used the Adam optimizer for training with a learning rate of 1e-4 and batch size of 8.

Table 2: Emotion Dataset Analysis

Emotion	No of images	Percentage of labels
Angry	3993	13.85%
Disgust	436	1.51%
Fear	4103	14.24%
Happy	7164	24.86%
Neutral	4982	17.29%
Sadness	4938	17.13%
Surprised	3205	11.12%

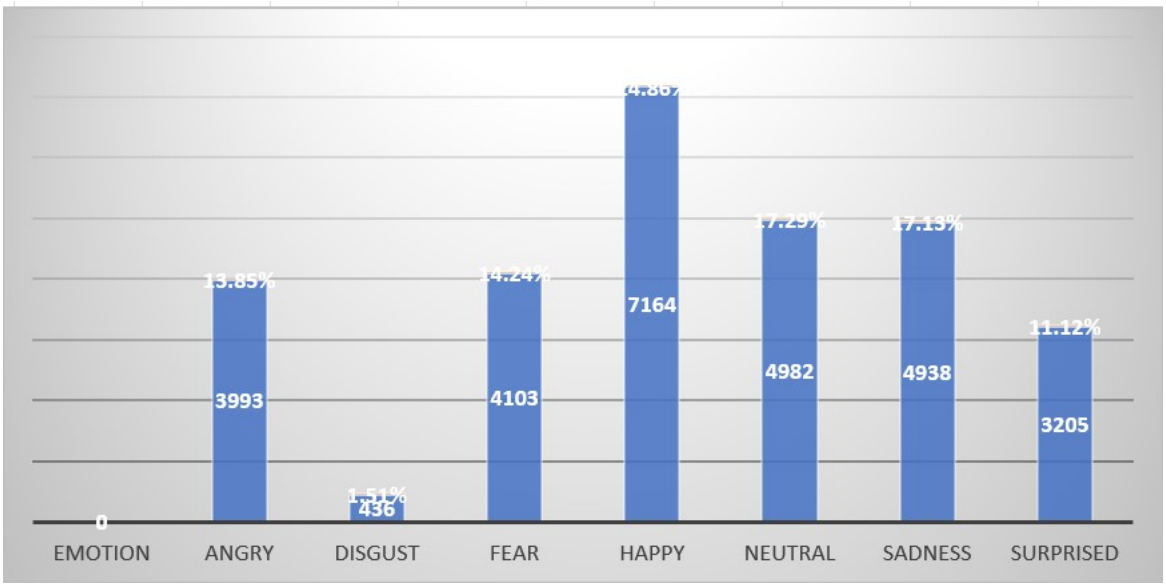

Fig 3: Data summary

Neural networks require large amounts of data for training and validation. The selection of high labelled data is directly responsible for the performance of the model. Thus, it requires high quality and quantity of data for training and validation. FERC-2013 dataset is employed for training the model, the dataset contains 48x48 grayscale images categorized into these emotions – happy, sad, angry, neutral, fearful, surprised and disgust [14]

Table 3: Table Showing Different Performance Metrics

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Mini-batch Loss	Base Learning Rate
1	150	00:00:48	21.88%	1.7765	1.0000e-04
2	350	00:01:53	41.41%	1.6193	1.0000e-04
3	500	00:02:43	42.19%	1.5446	1.0000e-04
4	700	00:03:43	42.19%	1.6481	1.0000e-04
5	900	00:04:45	43.75%	1.5707	1.0000e-04
6	1050	00:05:31	36.72%	1.5471	1.0000e-04
7	1250	00:06:31	52.34%	1.4098	1.0000e-04
8	1400	00:07:16	50.00%	1.3118	1.0000e-04
9	1600	00:08:16	41.41%	1.5294	1.0000e-04
10	1800	00:09:17	43.75%	1.4041	1.0000e-04

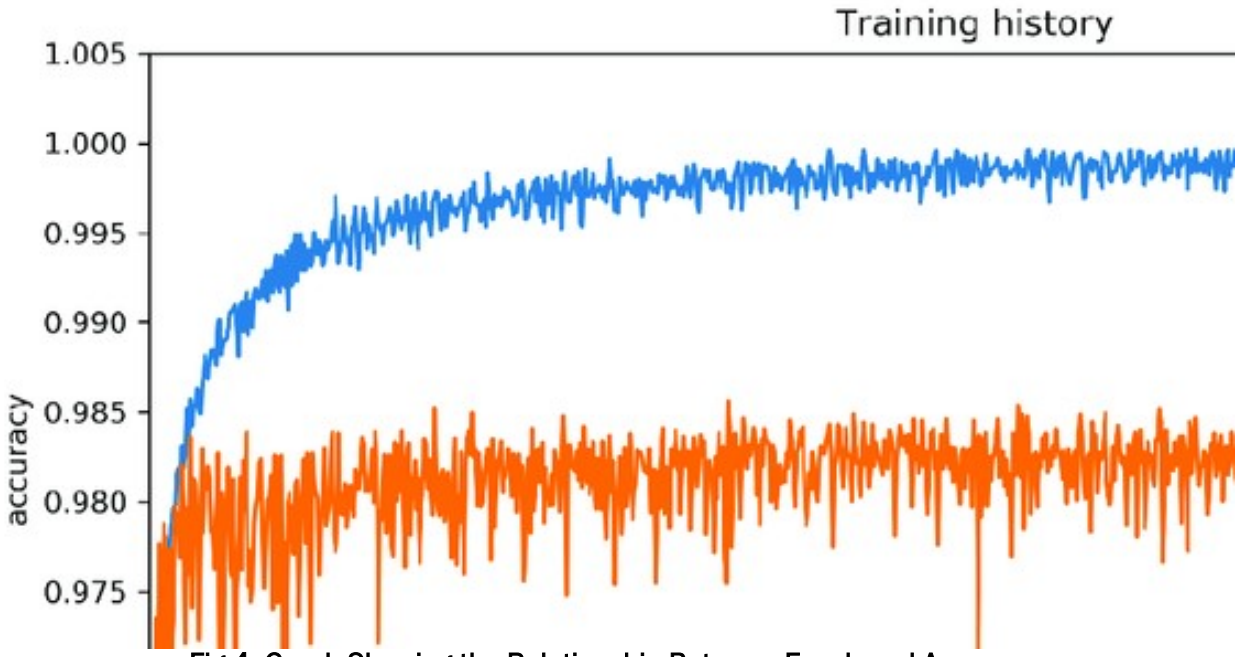

Fig 4: Graph Showing the Relationship Between Epoch and Accuracy

The model obtains accuracies of 44.28% and 39.58% on the validation and testing sets, respectively. The model performed well on the anger, happy, fear, sad, neutral and surprise emotions. In contrast, the disgust category was difficult to recognize. It may be caused by the imbalanced distribution of the data. From the data distribution shown in Table I, the label number ratios form these category is less than 20% in the entire dataset (disgust (1.51%), while the anger, happiness, sadness and surprise emotions categories compose most of the dataset.
A comparison of the various hyper parameters that were tuned can be seen in Table 4..

From table 4. shown above, we can deduce that the higher the epoch, the higher the batch size and the higher the number of iterations which improves the accuracy of the classified emotion in the facial emotion detection system. Fig. 4.1 shows the graphical illustration on the relationship between Epoch and the accuracy of results likewise table 4. it is noted that the higher the epoch, the more accurate the detection of the emotion on the system

Understandably happiness is very easy to determine as a direct result of the number of sample data present. Interestingly the emotion of surprise reached nearly the same accuracy. The other emotions had lower but similar accuracies. To put this system to test, a facial image is inputted into the system, processed. It was found that the model managed to predict almost all instances of happiness and most instances of surprise. It correctly predicted sadness and neutrality about half the time but it rarely predicted the other emotions correctly. The emotions of anger and fear in particular tended to mix while disgust was almost never predicted. It is also to be noted that in most cases of a wrong prediction, the second most likely prediction was often the right one.

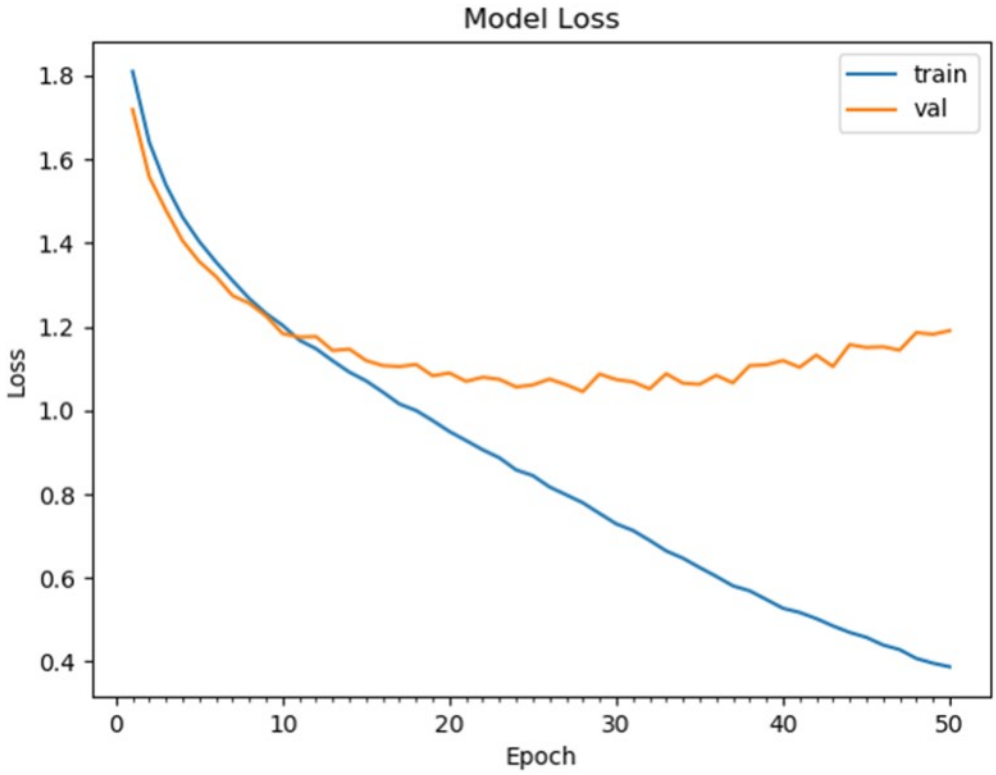

Fig 5: Loss Curve

ResNet, which stands for residual networks, is widely used CNN model for computer vision tasks. This model was awarded the ImageNet Challenge in 2017. This model has 50 convolutional layers. ResNet solved the problem of vanishing gradients which resulted for easier training. We are using the ResNet50 variation of this model for our application. We use pretrained “ImageNet” weights loaded into the model before our training and fine tuning. We add 2 more dense layers on top of this model. The first dense layer is equipped with a L2 class layer weight regularizer.

This is added to apply penalties and regularize layer output. Lastly, we add a 7 neuron SoftMax activation output layer representing our 7 emotion classes. Dropout layers with a factor of 0.2 are added in-between the dense layers as well to prevent overfitting. All the layers in this model are set to non-trainable except for last four layers for initial training. Same image augmentation techniques are used on the images that was discussed in the previous model consisting of rescaling, width shifting, height shifting, zooming and horizontal flip. The ResNet50 model is designed to take RGB images as input. It takes 3 layers of RGB as input but our dataset images are grayscale having only one grayscale layer

Table5: Tale Showing the Accuracy results

Emotions	No of images	No of corrected emotions	Accuracy
Angry	798	750	94%
Disgust	87	30	34%
Fear	820	614	75%
Happy	1432	1420	99%
Neutral	996	910	91%
Sad	988	790	80%
Surprise	641	622	97%

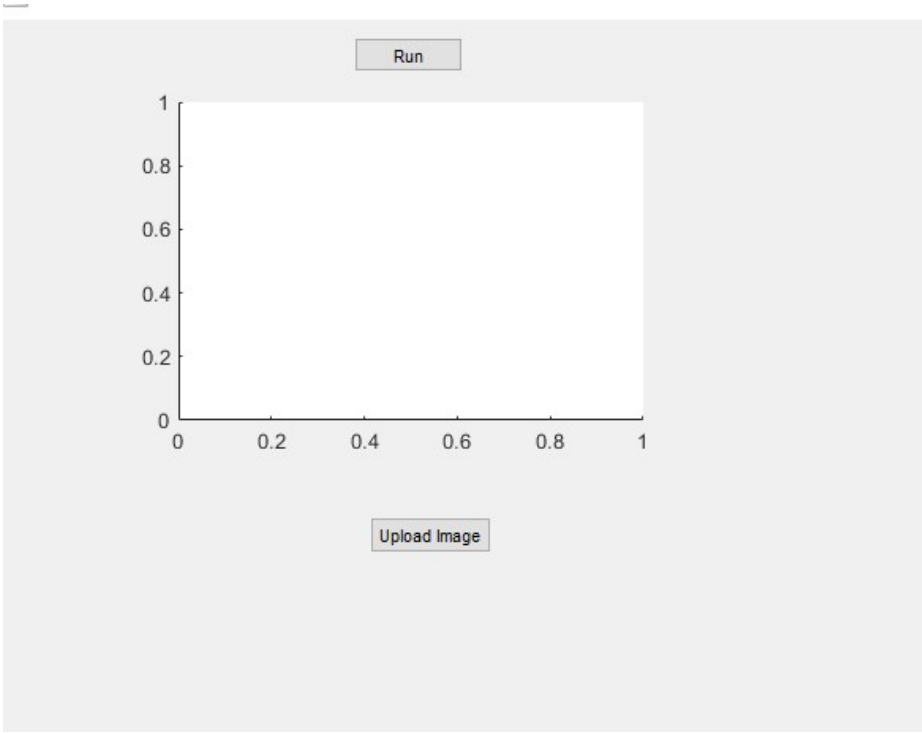

Figure 6: Graphical User Interface showing training and testing phase

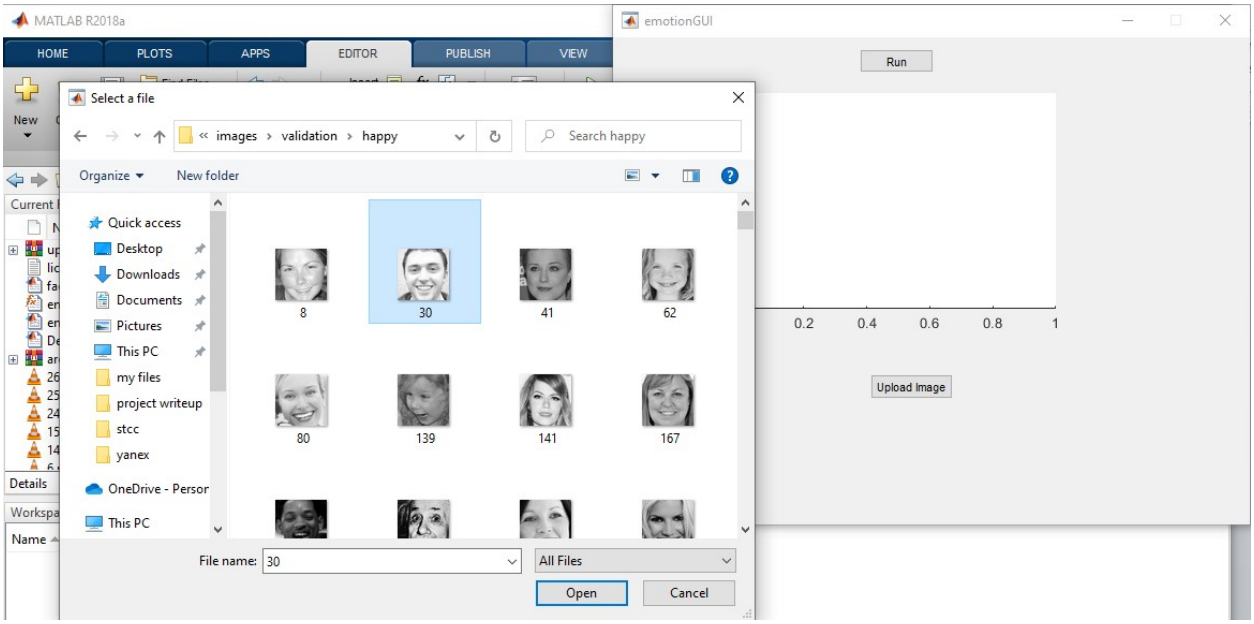

Figure 7: Graphical User Interface Showing The Library Directory For Uploading Image

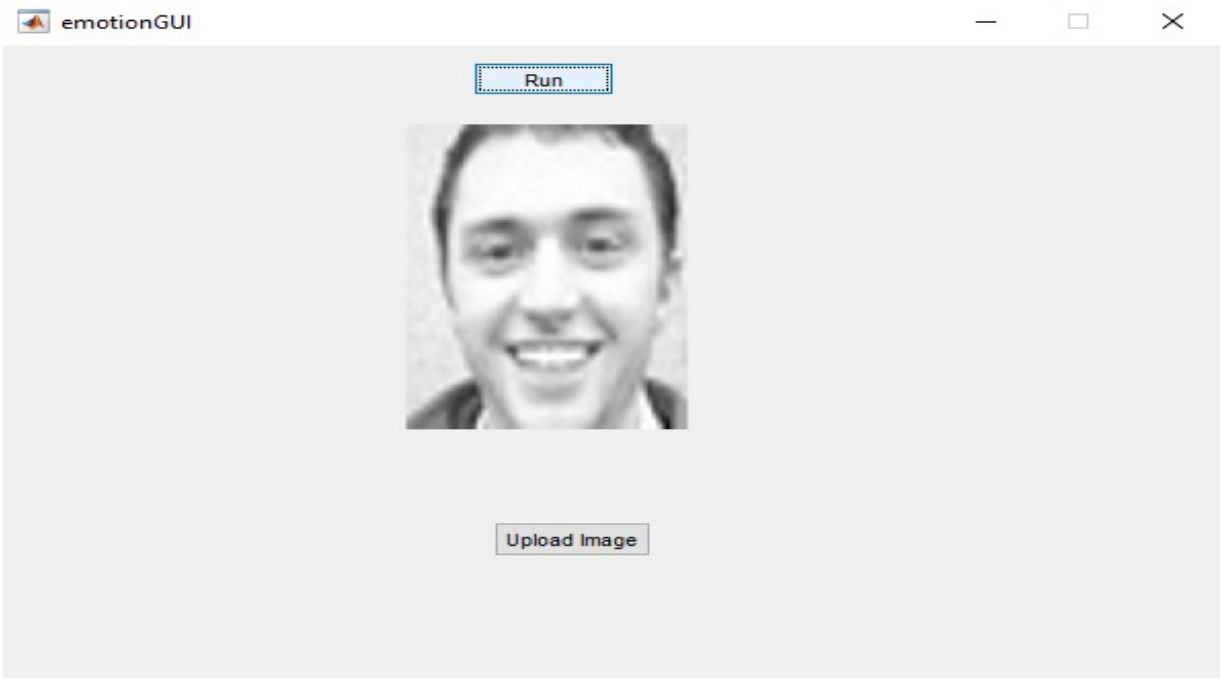

Figure 8: Graphical User Interface of Input Image

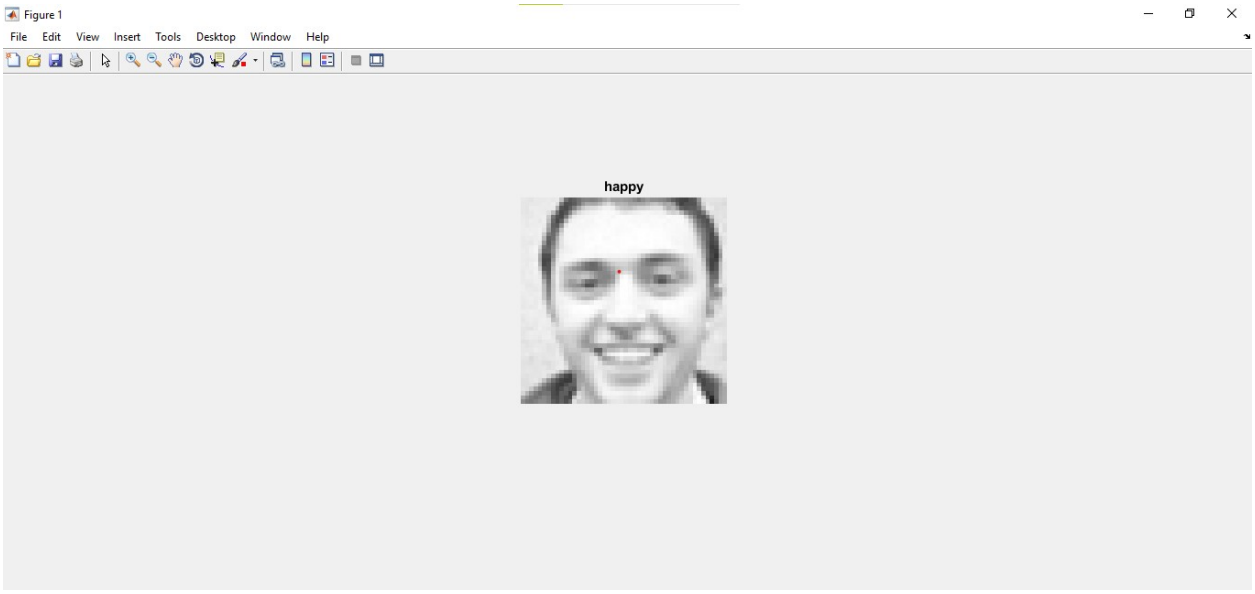

Figure 9: Graphical User Interface showing output (Happy)

The Input Image Graphical User Interface (GUI) plays a crucial role in enabling user interaction with the system. This GUI serves as the interface through which users can provide images for emotion detection. It is a user-friendly image upload feature that allows users to select or drag-and-drop images for analysis. This feature should support common image file formats such as JPEG, PNG, and GIF

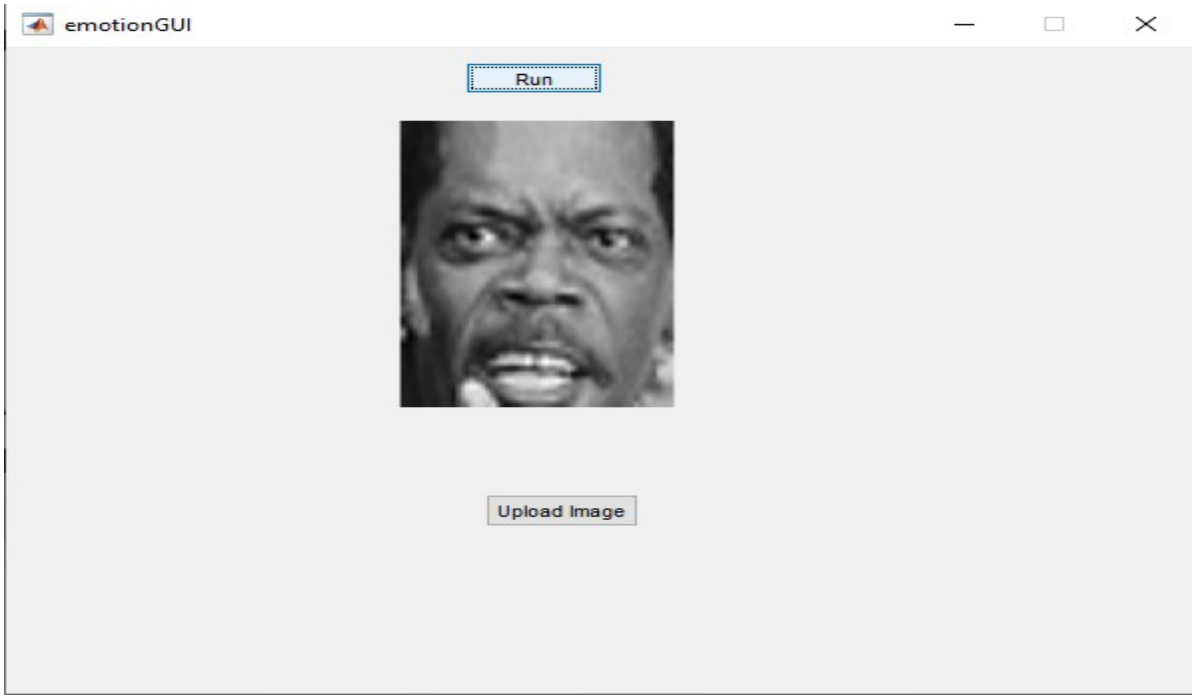

Figure 10: Graphical User Interface of input image

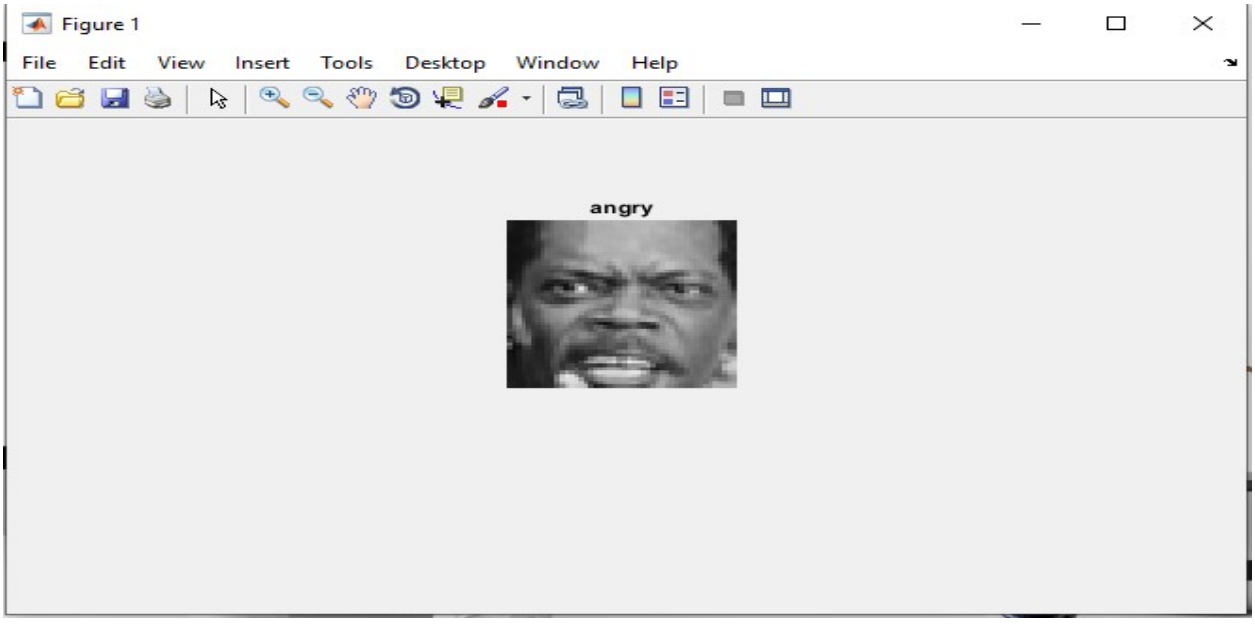

Figure 11: Graphical User Interface showing output (Angry)

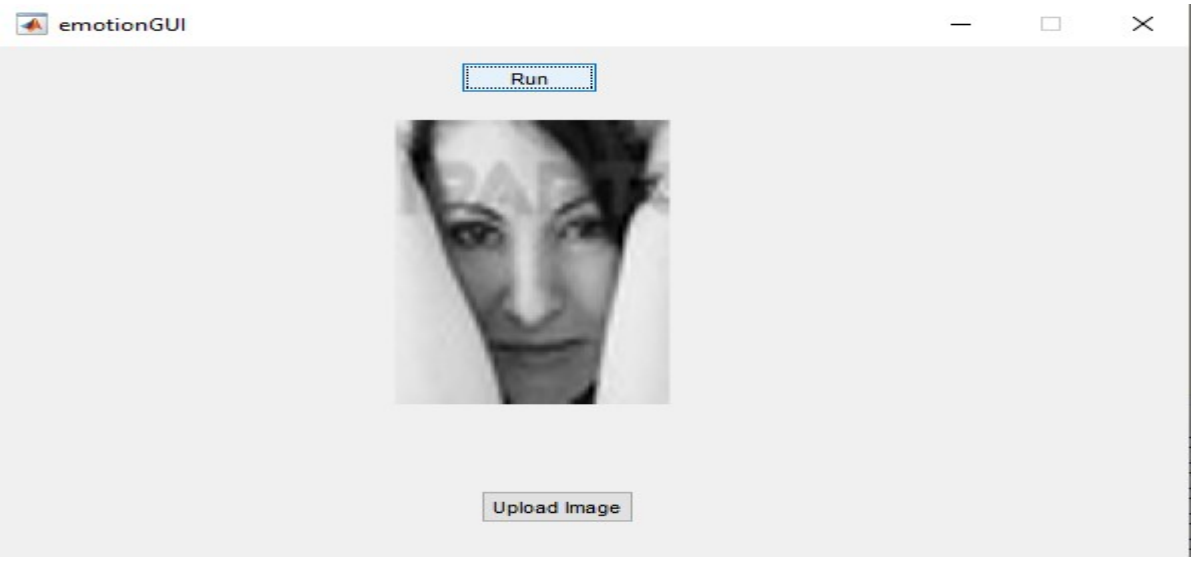

Figure 12: Graphical User Interface of input image (Fear)

The image displaying facial expressions and cues typically associated with fear or anxiety. This include widened eyes, raised eyebrows, and a tense or startled facial appearance. The system has identified and labeled the primary emotion portrayed in the image as fear, indicating that the person's facial expression conveys a sense of apprehension or distress

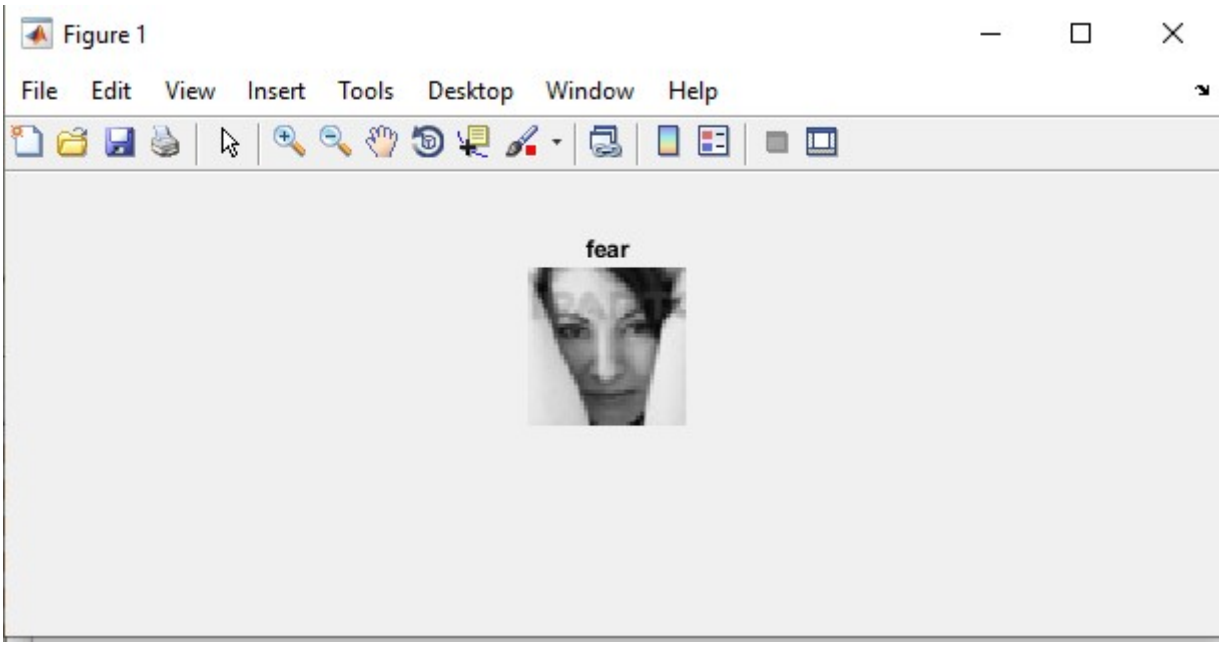

Figure 13: Graphical User Interface of output image (fear)

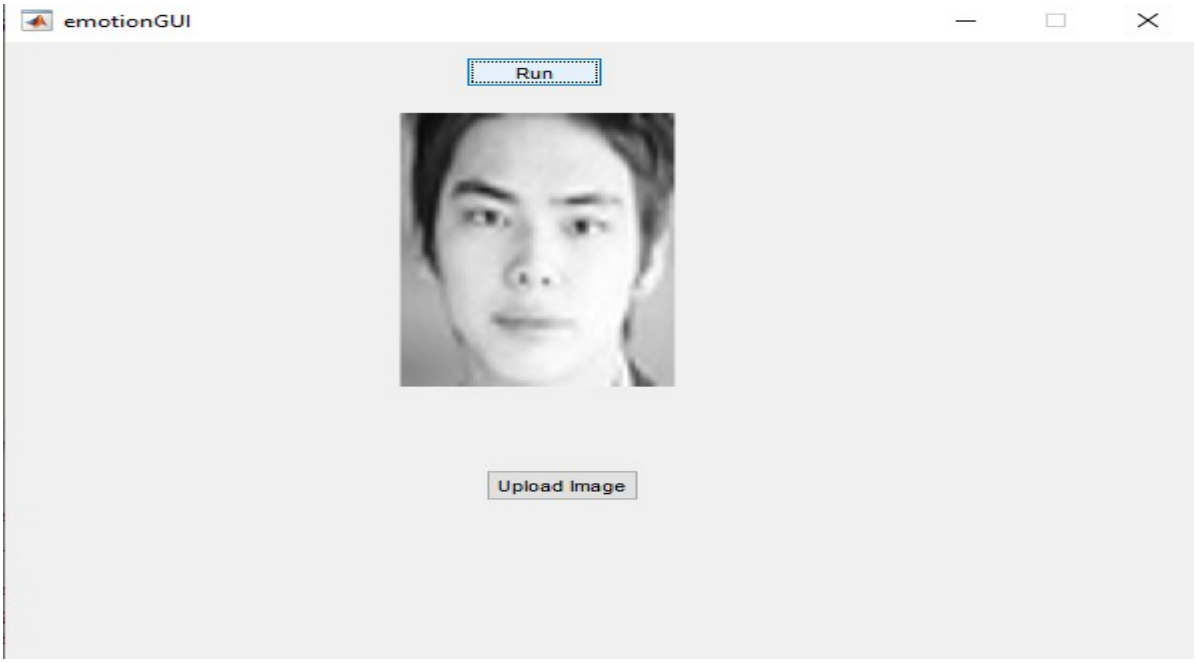

Figure 14: Graphical User Interface of input image (Neutral)

The image displaying a facial expression it appears calm, balanced, and devoid of strong emotional cues. It signifies a lack of pronounced emotions such as happiness, sadness, anger, or surprise. The system identified and labeled the primary emotion portrayed in the image as neutral, indicating a relatively emotionless or neutral facial expression.

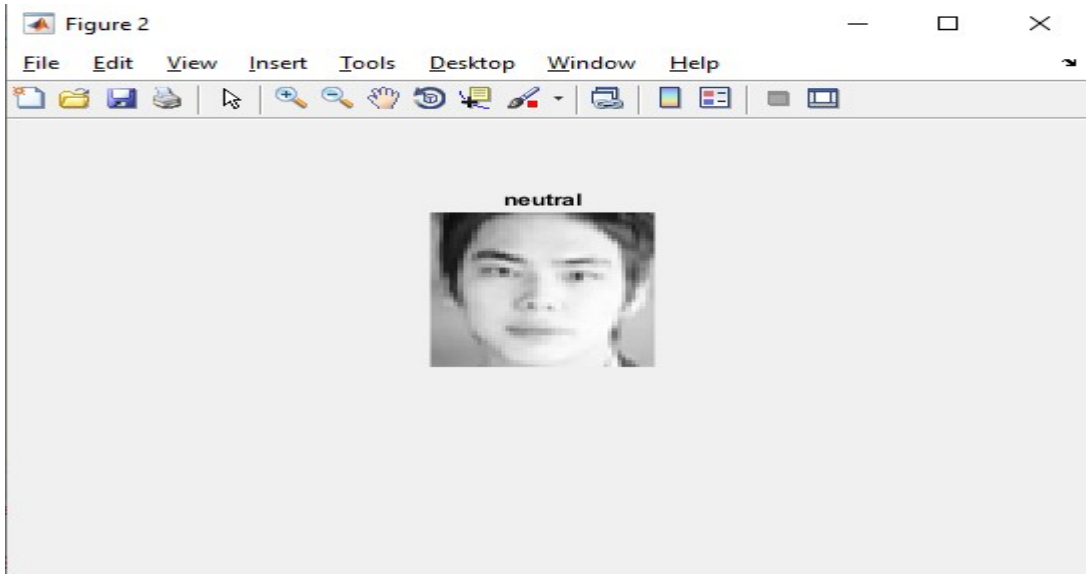

Figure 15: Graphical User Interface of output image (Neutral)

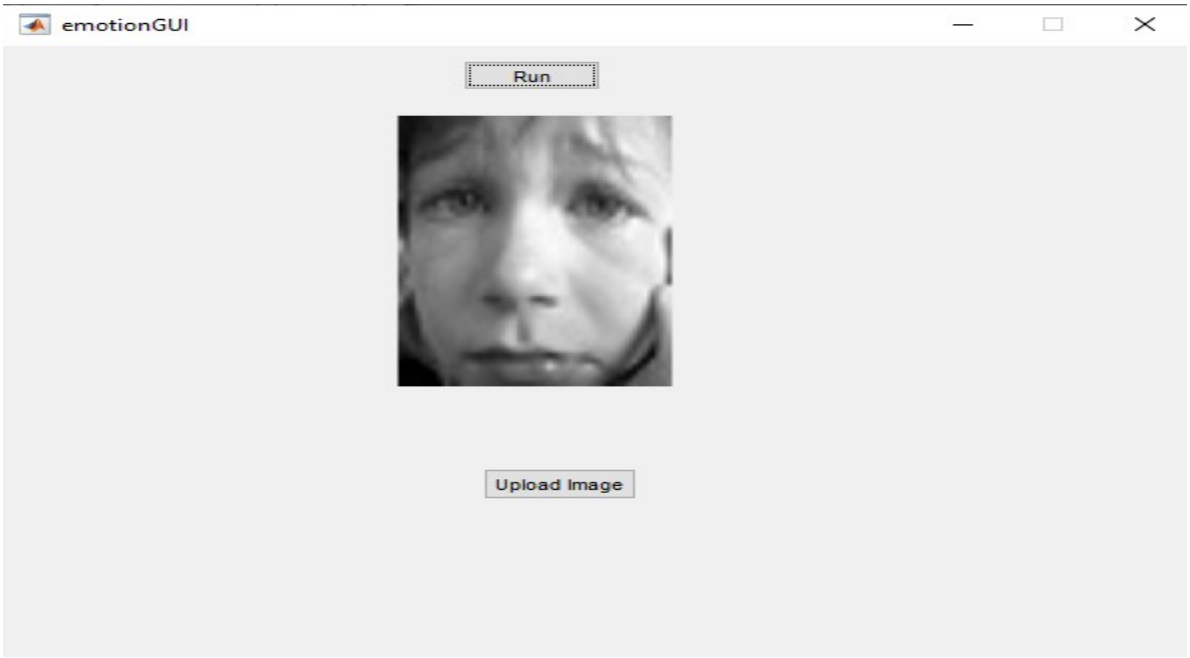

Figure 16: Graphical User Interface of input image (sad)

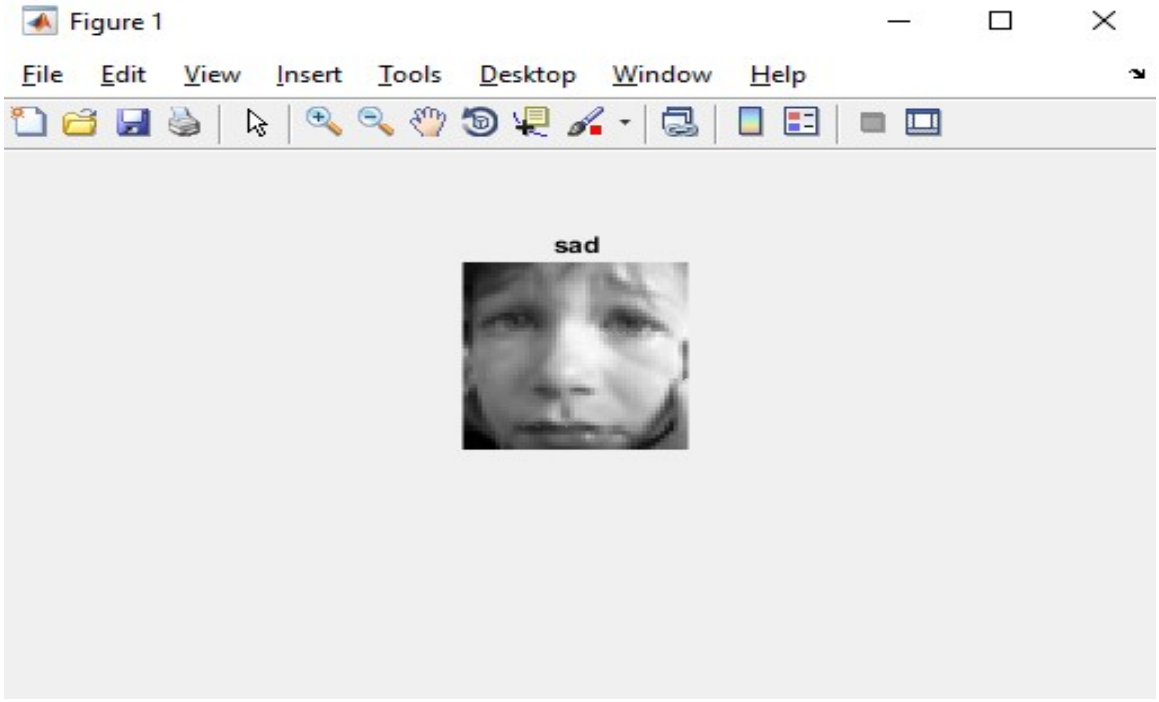

Figure 17: Graphical User Interface of output image (sad)

The image displaying facial expressions and cues typically associated with sadness. This is including a downturned mouth, drooping eyes, or other signs of unhappiness. The system has identified and labeled the primary emotion portrayed in the image as Sad.

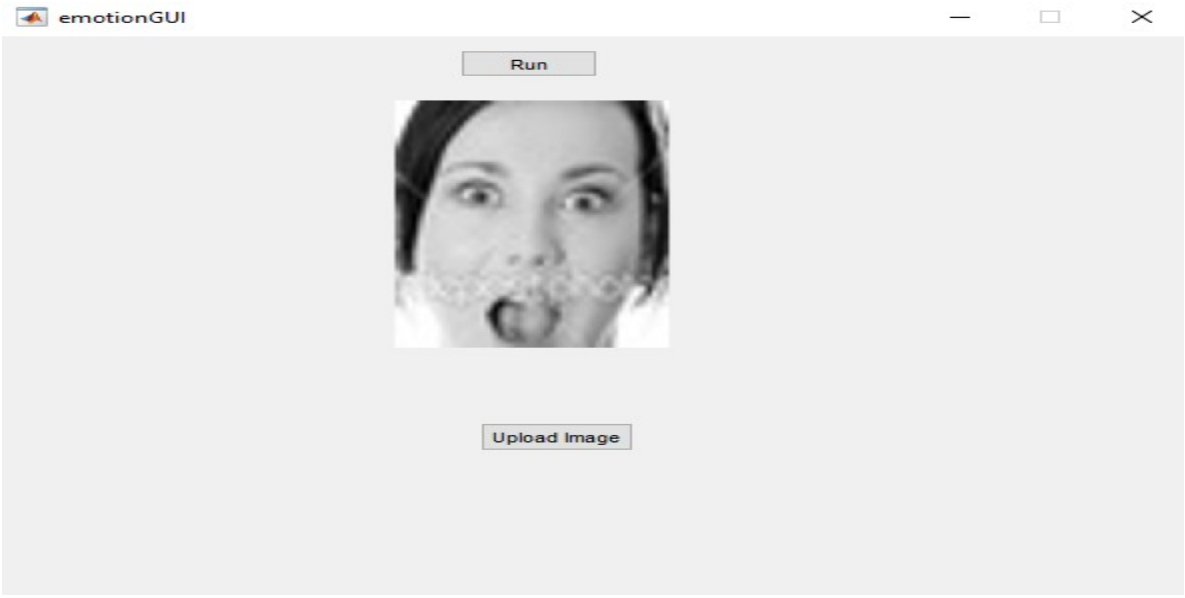

Figure 18: Graphical User Interface of input image (surprise)

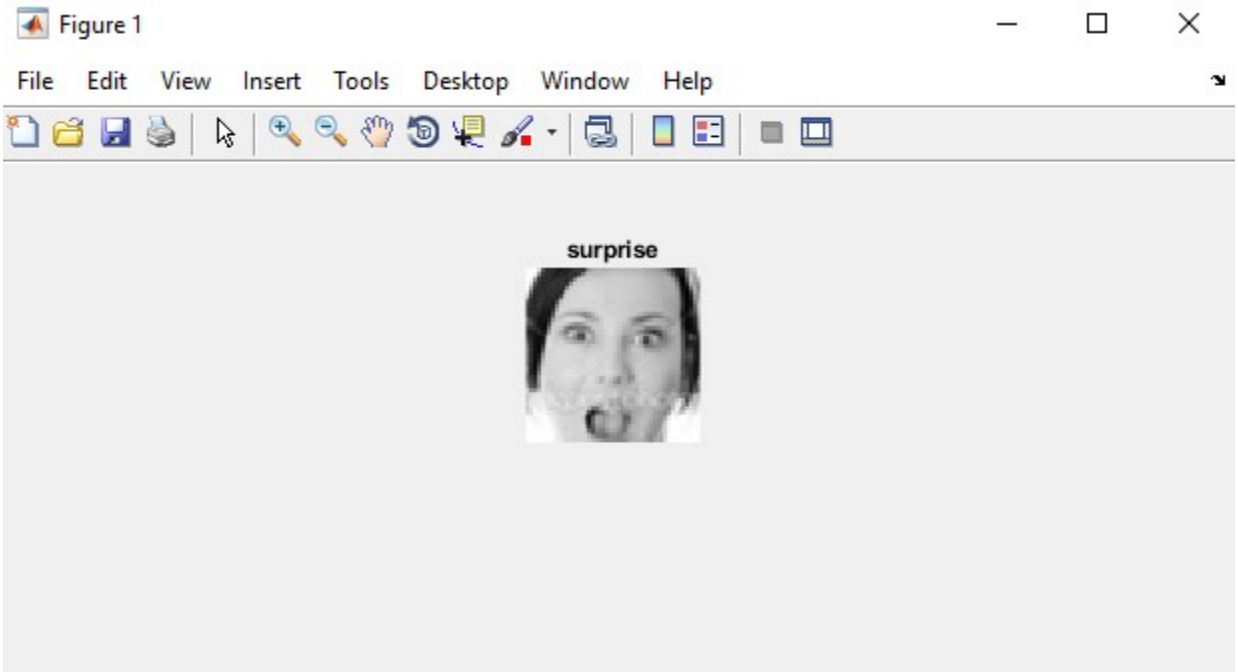

Figure 19: Graphical User Interface of output image (surprise)

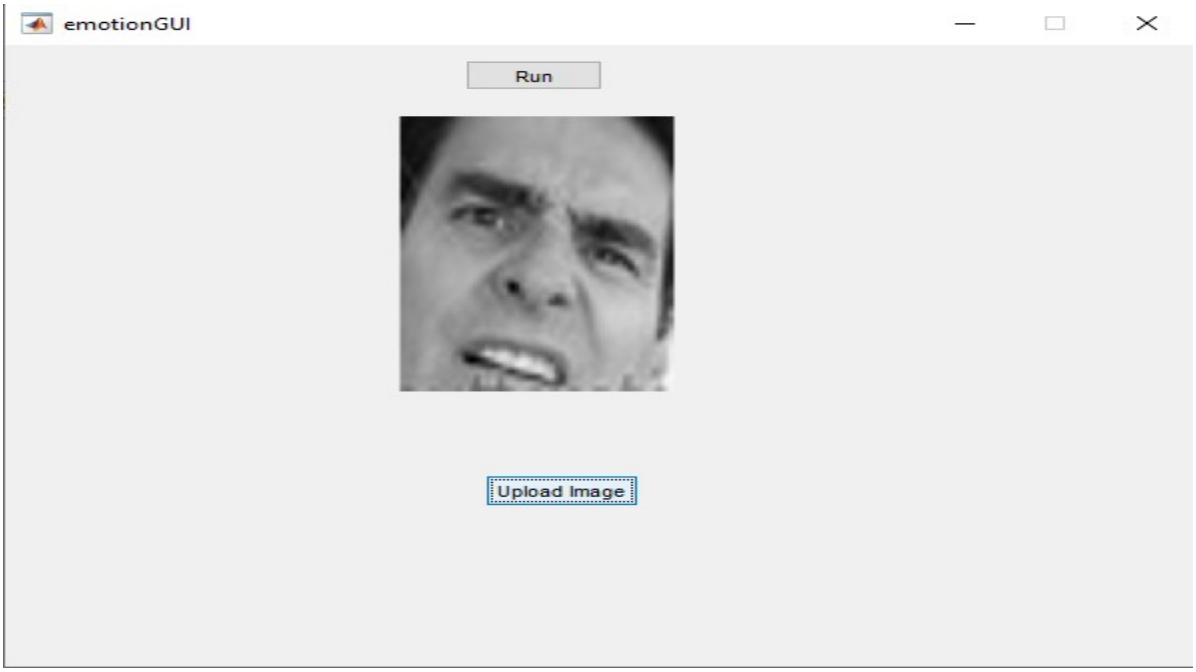

Figure 20: Graphical User Interface of input image (disgust)

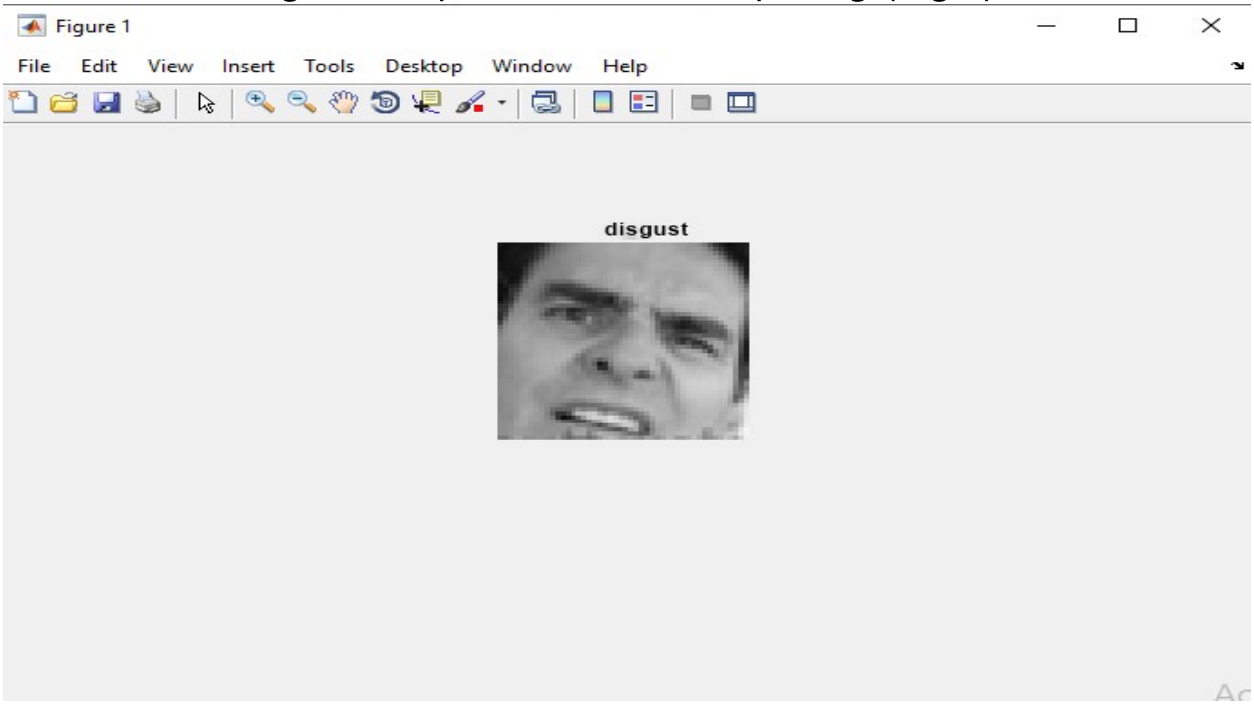

Figure 21: Graphical User Interface of output image (disgust)

Source Code

```
function varargout = emotionGUI(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @emotionGUI_OpeningFcn, ...
                   'gui_OutputFcn',  @emotionGUI_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end
 if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
 function emotionGUI_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;
guidata(hObject, handles);
 function varargout = emotionGUI_OutputFcn(hObject, eventdata, handles)
varargout{1} = handles.output;

function pushbutton1_Callback(hObject, eventdata, handles)
global myImage
startingFolder = 'C:\Program Files\MATLAB';
if ~exist(startingFolder, 'dir')
  % If that folder doesn't exist, just start in the current folder.
  startingFolder = pwd;
end
defaultFileName = fullfile(startingFolder, '*.*');
[baseFileName, folder] = uigetfile(defaultFileName, 'Select a file');
if baseFileName == 0
```

5. CONCLUSION AND RECOMMENDATION

An offline face recognition system with GUI is developed. The MEX function is successfully used to call the C algorithm into MATLAB environment. The system is developed based on four-layers CNN. However, only 2-layers of CNN are involved in retraining new incoming subjects into the system. This face recognition system has a faster retraining process, compared to retraining with 4-layers CNN. The epoch used for training in this system is less than 15 epochs. The accuracy of the system is 89.5% for all the seven emotions. However, MATLAB platform is not suitable for this system as there was result degradation where the average time to train was not consistent. For future work, this face recognition system could be developed in other platform; such as C, and could be extended as a real-time system. The system should have the ability to capture new images and saved into the image database.

However, deep-learning-based FER approaches still have a number of limitations, including the need for large-scale datasets, massive computing power, and large amounts of memory, and are time consuming for both the training and testing phases. it's imperative to acknowledge the ongoing constraints within deep learning-based Facial Expression Recognition (FER) approaches. These limitations span the demand for extensive datasets, substantial computational resources, substantial memory resources, and protracted time commitments for both training and testing phases. Conclusively, this project embodies an intersection of technological innovation and human emotion, yielding a functional face recognition system while laying the groundwork for continued advancements in the realm of FER.[20]

6. RECOMMENDATION

With regard to the performance of the developed technique; CNN based facial emotion detection system can be used to enhance education challenges in real world scenario.
It is recommended that:

- Other emotion detection techniques i.e, speech should be compared with the facial looks in other to determine its computational efficiency on emotion detection systems.
- A computer system with higher configurations and capability should be employed in other to handle more datasets because test-running the system with large dataset took a longer time to process.
- It also recommends motivational content for users, to overcome depression and sadness and stay delighted.
- In future, platform like this will be able to capture and detect facial expressions through small videos as input.[18]

REFERENCE

Ahsan, R., Hossain, M. S., & Rahman, M. M. (2019). Facial Expression Recognition Using Deep Convolutional Neural Network. In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

Akash Saravanan, Gurudutt Perichetla, K.S.Gayathri. (2019). Facial Emotion Recognition using Convolutional Neural Networks. arXiv:1910.05602v1 [cs.CV] 12 Oct 2019

Chahak Gautam, Seeja K.R. (2023) Facial emotion recognition using Handcrafted features and CNN. International Conference on Machine Learning and Data Engineering Procedia Computer Science 218 (2023) 1295–1303

Huang, W., Tang, Y., Chen, L., & He, Z. (2019). Facial Expression Recognition Based on Multi-Level Convolutional Neural Network. In 2019 Chinese Control and Decision Conference (CCDC) (pp. 614-619). IEEE.

Li, Z., & Chen, X. (2020). Facial Expression Recognition Based on Convolutional Neural Network and Capsule Network. In 2020 2nd International Conference on Computer Science and Software Engineering (CSSE) (pp. 118-121). IEEE

Lin, X., Zhou, H., & Huang, X. (2019). Facial Expression Recognition Based on Convolutional Neural Networks with Squeeze-and-Excitation Blocks. In 2019 Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 328-339). Springer

Liu, Y., Qin, T., Chen, M., Han, Y., & Wang, Y. (2020). Facial Emotion Recognition Based on Convolutional Neural Network with Spatial Pyramid Pooling. In 2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS) (pp. 226-231). IEEE

Makarand Madhavi, Isha Gujar, Viraj Jadhao, and Reshma Gulwani. (2022). Facial Emotion Classifier using Convolutional Neural Networks for Reaction Review. ITM Web of Conferences 44, 03055 (2022)

MalyalaDivya, R Obula Konda Reddy, C Raghavendra. (2019). Effective Facial Emotion Recognition using Convolutional Neural Network Algorithm. ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Neha Deshpande, Fabrizio Nunnari, Eleftherios Avramidis. (2022). Fine-tuning of Convolutional Neural Networks for the Recognition of Facial Expressions in Sign Language Video Samples. Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 29–38

Pooja Bagane, Shaasvata Vishal, Rohit Raj, Tanushree Ganorkar, Riya. (2022). Facial Emotion Detection using Convolutional Neural Network International Journal of Advanced Computer Science and Applications, Vol. 13, No. 11, 2022

Qi, Y., Qiao, Y., Gao, Y., & Zhang, T. (2020). Facial Expression Recognition Based on Modified Convolutional Neural Network. In 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 95-98). IEEE.

Sun, X., Li, Y., Huang, Z., Zhang, X., & Hu, W. (2020). Facial Emotion Recognition Based on Convolutional Neural Network and Ensemble Learning. In 2020 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) (pp. 191-196). IEEE.

Tao, J., Liu, Y., Sun, X., & Wu, Y. (2020). Facial Expression Recognition Based on Deep Convolutional Neural Network with Local Binary Pattern Features. In 2020 International Conference on Communication Technology and Application (ICCTA) (pp. 39-43). IEEE.

Wang, B., Cheng, H., Fu, Y., & Gu, Q. (2021). Facial Expression Recognition Based on Deep Convolutional Neural Networks. In 2021 4th International Conference on Electronic Information Technology and Industrial Development (ICEITID) (pp. 54-57). IEEE.

Xiang, J, Lin, Y, Zhang, Z, & Wei, S. (2020). Facial Expression Recognition Based on Deep Convolutional Neural Network and Class Activation Map. In 2020 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) (pp. 644-648). IEEE.