Computers, Materials & Continua
DOI: 10.32604/cmc.2022.029101
Article

Tech Science Press

Emotion Recognition from Occluded Facial Images Using Deep Ensemble Model

Zia Ullah[1], Muhammad Ismail Mohmand[1], Sadaqat ur Rehman[2,*], Muhammad Zubair[3], Maha Driss[4], Wadii Boulila[5], Rayan Sheikh[2] and Ibrahim Alwawi[6]

[1]Department of Computer Science, The Brains Institute, Peshawar, 25000, Pakistan
[2]School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, UK
[3]Department of Neurosciences, KU Leuven Medical School, Leuven, 3000, Belgium
[4]Security Engineering Laboratory, CCIS, Prince Sultan University, Riyadh, 12435, Saudi Arabia
[5]Robotics and Internet of Things Lab, Prince Sultan University, Riyadh, 12435, Saudi Arabia
[6]Department of Computer Science, Robert Gordon University, Aberdeen, UK
*Corresponding Author: Sadaqat ur Rehman. Email: sadaqatur.rehman@abdn.ac.uk
Received: 25 February 2022; Accepted: 24 May 2022

Abstract: Facial expression recognition has been a hot topic for decades, but high intraclass variation makes it challenging. To overcome intraclass variation for visual recognition, we introduce a novel fusion methodology, in which the proposed model first extract features followed by feature fusion. Specifically, RestNet-50, VGG-19, and Inception-V3 is used to ensure feature learning followed by feature fusion. Finally, the three feature extraction models are utilized using Ensemble Learning techniques for final expression classification. The representation learnt by the proposed methodology is robust to occlusions and pose variations and offers promising accuracy. To evaluate the efficiency of the proposed model, we use two wild benchmark datasets Real-world Affective Faces Database (RAF-DB) and AffectNet for facial expression recognition. The proposed model classifies the emotions into seven different categories namely: happiness, anger, fear, disgust, sadness, surprise, and neutral. Furthermore, the performance of the proposed model is also compared with other algorithms focusing on the analysis of computational cost, convergence and accuracy based on a standard problem specific to classification applications.

Keywords: Ensemble learning; emotion recognition; feature fusion; occlusion

1 Introduction

As the major technique of processing for nonverbal intentions, Facial Expression Recognition (FER) is a vital and vast branch of computer vision and machine learning, as well as one of the symmetry subject areas. Emotions are unavoidable in human interactions. They can appear in various ways and may not be visible to the naked eye. Therefore, we can use different tools to aid us in the detection and recognition of them. Human emotion recognition is becoming more popular in a variety of domains, including medicine [1,2], human-machine interfaces [3], urban sound perception [4], animation [5], diagnosis of autism spectrum disorder (ASD) in kids [6], and security [7,8] but the recognition is not limited to these fields only. Several features that include EEG [9], facial expressions [5,10,11], text [12], and speech are [7,13] used for the recognition of emotions. Face expression features are one of the most famous methods among other methods of human recognition due to many reasons. Some of the reasons are, i.e., (1) they are noticeable and visible, (2) large face datasets can be collected easily by face expression features (3) they also contain many features for the recognition of emotions [5,14,15]. By using deep learning, particularly CNN-based learnable image features [16] can also be computed, learnt, and extracted to recognize facial expressions [17,18].

FER research will increasingly concentrate on in-the-wild spontaneous expressions as Deep Learning technology advances and the need for applications grows in the age of big data. It is needful to propose new FER solutions in complex environments, such as occlusion, multi-view, and multi-objective. It is extremely important to give the classifier with the most relevant data under ideal conditions to obtain a proper facial expression classification. In traditional FER techniques the first stage is to pre-process the input image to accomplish this. Face detection is also a common pre-processing step in most peer-reviewed papers. However, many facial expressions can be cued from various regions of the human face, the region can be nose, mouth, cheeks, forehead, and eyes, while ears and hairs play a minor role in detecting facial expressions [19]. As more expressions can be observed by mouth and eyes, the computer vision deep learning model should emphasize these parts and overlook the other details of the face.

The following are significant contributions in to the proposed work:

- A new features extraction block is developed, extracting robust invariant characteristics from facial images using three state-of-the-art CNN models. The features extracted, through the application of using the three models, are concatenated using a feature fusing method.
- Three deep learning classifiers are selected and trained to develop an accurate facial expression predictive model. The three classifiers' output is combined using an ensemble classification approach by applying majority voting to upgrade the facial expression recognition accuracy.
- The results obtained from pre-trained models for feature extraction and ensemble classification approach techniques are compared with other most presented methods.
- The experiments are carried out on two publicly available datasets, and it was discovered that the suggested technique performs better in recognizing seven basic facial expressions.

2 Related Work

Zelier et al. [20] present a new visualization technique to understand the working of intermediate feature layers along with the classifier's operation. Paul Ekman et al. [21] identified the six primary emotions, i.e., As the most typical study in this field of emotion recognition, the emotion of pleasure, fear, hate, sorrow, disgust, and surprise (except neutral). Ekman et al. later used this concept to create Facial Action Coding System (FACS) [22], which became the gold standard for emotion recognition research. Neutral was later added to most human emotion recognition datasets; which results in seven fundamental human emotions.

An approach of two-step machine learning was used in early work on emotion recognition. The first stage involves extracting characteristics from the image, and the second involves applying a classifier to detect emotions. Some of the most common use manual features for the recognition of facial expressions are Gabor wavelets [23], Haar features [24], Texture features i.e., Local Binary Pattern (LBP) [25], and Edge Histogram Descriptor [26,27]. The best sentiment is then assigned to the image by the classifier. These methods appear to work well on more specific datasets, but as more challenging datasets (with more intra-class variation) become available, their limitations become apparent. Figs. 1 and 2 had an image that we are referring to the reader, showing only the parts of the face or the spectacles or the hand covered, to better appreciate some of the issues that images can confront.

Based on deep education, various organizations have achieved huge successes in neural networks and in deep learning, vision difficulties and image categorization, and have built facial expression recognition (FERs). Khorrami et al. [17] demonstrated that CNN could achieve a higher accuracy level for emotion recognition and on extended used zero-bias CNN's Toronto Face Dataset (TFD) and Cohn-Kanade dataset (CK+) for the achievement of state-of-the-art results. For modelling the expression on humans' faces, Clavel et al. [5] used deep learning to train a network and the other to map human images to animated faces to create a novel model for human facial emotion recognition of animated characters. Mollahosseini has suggested FER neural network with maximum layer of pooling, two convolution layers, and four "initial" layers or subnetworks [10]. Liu [13] employs one recurring network to combine the removal and classification of features, emphasizing the need for input from the two components. The Boosted Deep Belief Network (BDBN) was used to deliver state of the art CK+ and JAFFE accuracy.

On noisy labels of authentic images obtained through crowdsourcing, Barsoom et al. [25] used a deep CNN. They employed ten taggers to re-enact each image to obtain acceptable precision, by 10 tags in their dataset they used many costing functions for their Deep CNN. Han et al. [26] to enhance spontaneous identification of face expression by increasing the discriminative neurons that outperformed their best at the time technique introduced Incremental Boosting CNN (IB-CNN). Meng [27] developed an identity-aware CNN that decreases variance in expression-related information while learning identity by employing identity and expression sensitive contrast loss. Finally, Fernandez et al. devised a network design called an end-to-end network architecture with a focus model [28].

Want et al. [29] introduced a simple and effective self-repair technique to eliminate uncertainty and avoid uncertain face images (due to labelling noise) from overfitting the deeper Self Cure Network (SCN). In two dimensions, SCN minimizes uncertainty: (1) by the employment of a self-attention mechanism to weight each sample of workouts on tiny batches with rank regularization; and (2) by a meticulous rebellion process to alter these samples in the lowest rank set. An algorithm was developed to recognize facial expressions by Wang et al. That is used for the occlusion changes and the resistance of pose in the real world [30]. To capture the importance of pose variant FER and facial regions in occlusion, they proposed a new network called Regional Attention Network (RAN). Some of the other works for the recognition of facial expressions are deep self-attention networks for the recognition of face emotion [31], multi-attention networks for the recognition of facial expression [32] and a new review on emotion recognition using facial appearance [33]. In [34], a lightweight CNN is developed that may effectively handle the problem of model overfitting caused by insufficient data while also utilizing transfer learning.

All the works that are mentioned above have improved emotion recognition significantly compared to previous work. Still, none of these works contains a simple method for identifying essential face regions to detect emotions. This research proposes a new framework based on a further attentiveness-coevolutionary neural network to concentrate on crucial facial areas to tackle this challenge.

3 Materials and Methods

3.1 Datasets and Pre-Processing

The Real-world Affective Faces Database (RAF-DB) [35] and AffectNet [36] are among the benchmark datasets used in this work for facial expression recognition. Let's take a short look at these databases before we get into the findings.

3.1.1 RAF-DB

RAF-DB is a publicly available dataset. The collection contains 30,000 pictures with a resolution of 48 by 48 pixels, the majority of which were taken in the field. Each image has been given its own label by nearly 40 annotators using crowdsourced annotation. Only images exhibiting basic emotions were employed in our study, with 12,271 images used as training data and 3,068 images used as test data. Six images are given in Fig. 1 from the dataset of RAF-DB.

Figure 1: For example, six images from the RAF-DB are shown

3.1.2 AffectNet

AffectNet is the world's biggest publicly available collection of huma facial expressions, arousal, and valence, which allow researchers to study automated facial expression identification in two separate emotion models. The version of the AffectNet dataset we used for our experiments contains 291,651 training samples and 4,000 tests samples. They manually annotated the images with eight different facial expressions Neutral, Happy, Surprise, Sad, Fear Disgust and Anger. These images are cropped and resized to 224 × 224 pixels. Fig. 2 showed four images from the dataset.

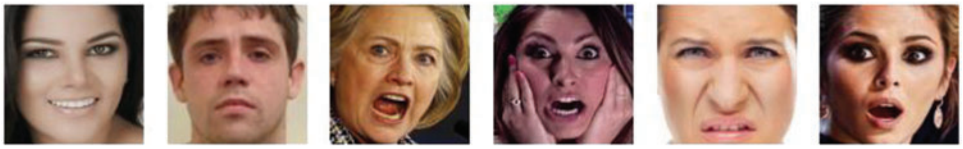


Figure 2: Six images from the AffectNet dataset as examples

3.2 Methodology

The underlying strategy of our study is briefly outlined in this part, followed by an explanation of DCNN models. Following that, we present the decision fusion strategies that we have implemented. As shown in Fig. 3, the basic concept of our study is identical to any traditional human Facial Expression Recognition technique. Direct training of deep networks on facial datasets is prone to overfitting.

The suggested methodology's workflow consists of four steps: feature extraction using three state-of-the-art CNN models, feature fusing, ensemble learning model, and finally applying the majority voting scheme to the output of three classifiers. Two benchmark datasets of seven classes each are loaded to the system, partitioning into two sub-data, i.e., trainset and validation. In the first step of the proposed methodology, three state-of-the-art CNN architectures are used to extract robust and non-invariant features from the RGB images, i.e., Inception-V3, VGG-19, and Resnet50 extract features from the training set image and validation set images. The second stage entails, feature fusion is introduced to fully describe the rich internal information of image features from each model. In the third stage for the emotion's recognition task, the deep ensemble learning model is developed using three state-of-the-art classifiers, i.e., CNN, Long short-term memory (LSTM), and Gated Recurrent Unit (GRU). In the third and final stages, a majority voting scheme is applied to the output of three classifiers. The output with maximum is selected as the final prediction of the proposed emotion recognition model. A detailed explanation of each stage in the proposed model is described in the following subsections.

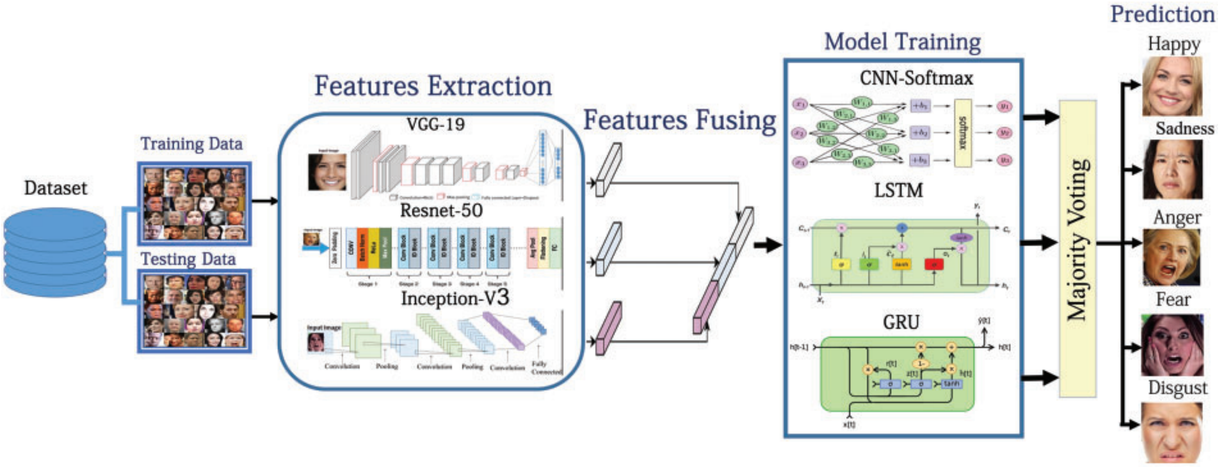


Figure 3: Our proposed facial expression recognition model

3.3 *Pre-Trained Convolutional Network Architectures*

Below is a quick description of the three state-of-the-art DCNN models for deep feature extraction that were chosen.

3.3.1 *VGG-19*

It is created by increasing the depth of the available CNN model to sixteen or nineteen, as illustrated in Fig. 4. The visual geometry group (VGG) proposed VGGNet architecture for ILSVRC 2014 and won the challenge [37]. VGG-19's architecture contains 144 million parameters, while VGG-16 has 138 million. There are 13 convolutional layers in the VGG-16, five max-pooling layers (22), and two fully-connected layers. ReLU activation is used in all convolution layers, while dropout regularization is used in fully connected layers. We retrieved deep features from the final fully-connected layers in our research.

3.3.2 *ResNet-50*

It is a deep convolutional network that solves the problem of vanishing gradients by identifying convolutional blocks. As the gradient is back-propagated via a deep network, it may become incredibly small. The identity block solves the problem of vanishing gradient by using shortcut connections, which are another possible way for the gradient to travel. ResNet with 50 convolutions in five stages would be used in our methodology, as shown in Fig. 5. Each stage featured a convolutional and an identity block, with three convolutions in each block, with 1 × 1, 3 × 3, and 1 × 1 filters, with the 1 × 1 kernel responsible for lowering and then increasing dimensions. For bottleneck architectures, parameter-free identity shortcuts are especially crucial. The model size along with time complexity are doubled when the identity shortcut is substituted with the projection, because the shortcut connects the two high-dimensional endpoints. As a result, identity shortcuts lead to more efficient bottleneck design models [38].

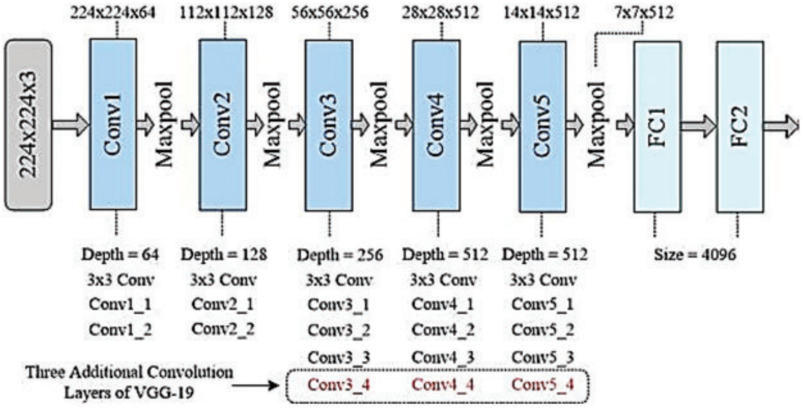
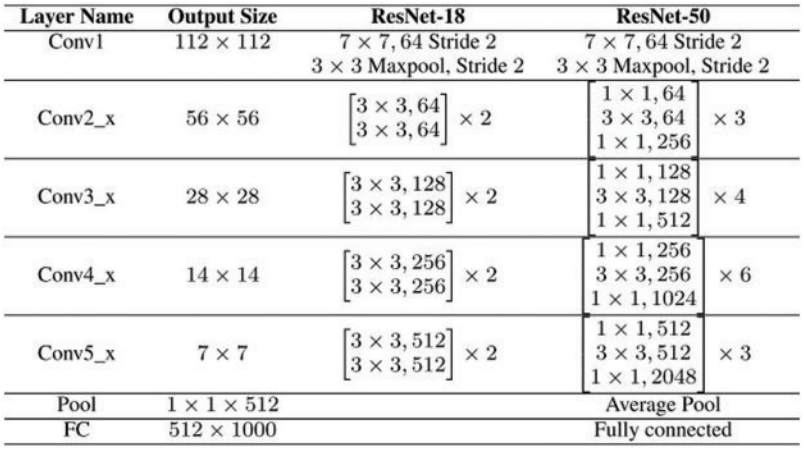


Figure 4: Architecture of VGG-16 and VGG-19



Layer Name	Output Size	ResNet-18	ResNet-50
Conv1	112 × 112	7 × 7, 64 Stride 2 3 × 3 Maxpool, Stride 2	7 × 7, 64 Stride 2 3 × 3 Maxpool, Stride 2
Conv2_x	56 × 56	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$
Conv3_x	28 × 28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$
Conv4_x	14 × 14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times 2$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 6$
Conv5_x	7 × 7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times 2$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$
Pool	1 × 1 × 512		Average Pool
FC	512 × 1000		Fully connected

Figure 5: Basic ResNet-18 and ResNet-50 architecture

3.3.3 Inception-V3

Fig. 6 shows Inception V3, it is a deep neural network having 42 layers which minimized emblematic bottlenecks. Inception V3 had five stem convolutional layers which consists of a type-A reduction block, three type-A Inception blocks, a reduction block, four type-B Inception blocks, two type-C Inception blocks, an average pooling layer, and lastly the fully connected network. Factorization was considered to minimize the size of a deep neural network. To minimize overfitting, several factorization modules were inserted in the convolutional layers to lower the size of the model. The performance of neural networks was better when convolutions did not drastically reduce the size of the input, causing information loss. Splitting 5 × 5 convolutions into two 3 × 3 convolutions was one factorization (type-A inception block). Furthermore, factorization of the $n \times n$ filter to a mixture of $1 \times n$ and $n \times 1$ asymmetric convolutions (type-B inception block) was discovered to significantly reduce computing costs. It has been discovered that while on early layers, this factorization does not function well, this performs admirably on middle grid sizes [39]. The high-dimensional representations were the last factorization considered, which involved replacing two of the 3 × 3 convolutions with asymmetric 1 × 3 and 3 × 1 convolutions.

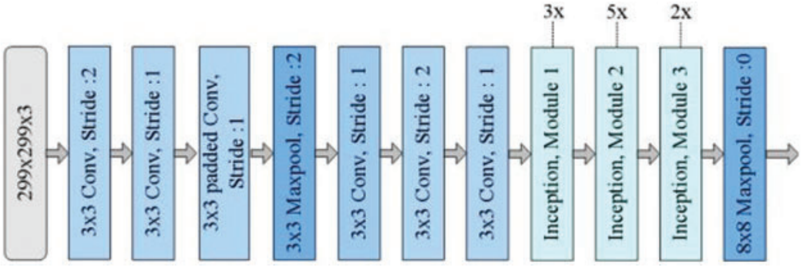


Figure 6: Basic inception-V3 architecture

3.3.4 Feature Fusing

Feature fusion allows us to fully describe the rich internal information of image features from each model, and after dimensionality reduction, we can obtain compact representations of integrated features, resulting in lower computational complexity and better face detection performance in an unconstrained environment [40]. After, all the feature sets are combined to generate a new one.

Let the three features retrieved from VGG19, ResNet-50, and Inception V3 be X, Y, and Z, respectively. Let $\mho$ be the pattern sample space, and φ be a randomly selected sample in $\mho$. In addition, α, β, and γ are the feature vectors of φ, where $\alpha \epsilon X$, $\beta \epsilon Y$ and $\gamma \epsilon Z$ respectively.

A combined feature's definition can be found in Eq. (1), where δ represent the serial combined feature. If the dimensions of α, β and γ are n_1, n_2 and $n_3$respectively, then the dimension of δ is given as $n_1 + n_2 + n_3$.

$$\delta = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \tag{1}$$

The complex vector given in Eq. (2) can be used to define the parallel feature fusion approach of φ, where j and k denote imaginary units. In case the dimensions of α, β, and γ are not equal, lower-dimensional features should be padded with zero's, such that before being joined, all of the features have the same dimension.

$$\delta = \alpha + j\beta + k\gamma \tag{2}$$

In the proposed technique, we adopt a weighted serial feature fusion methodology to integrate three feature vectors, which alters the serial feature fusion strategy. After normalization, the global, appearance, and texture feature vectors are denoted by the letters $f1, f2$, and $f3$, respectively. After then, the fusion feature F can be obtained, as shown in Eq. (3), where the weights of f_1, f_2 and f_3 are w_1, w_2 and w_3 respectively.

$$F = \begin{bmatrix} w_1 f_1 \\ w_2 f_2 \\ w_3 f_3 \end{bmatrix} \tag{3}$$

The single recognition rate of f_1, f_2 and f_3, which are indicated by A_1, A_2 and A_3, respectively, determines the values of weights w_1, w_2 and w_3. We calculate the values of w_1, w_2 and w_3 using Eq. (4) to Eq. (6).

$$w_1 = \frac{A_1}{A_1 + A_2 + A_3} \tag{4}$$

$$w_2 = \frac{A_2}{A_1 + A_2 + A_3} \tag{5}$$

$$w_3 = \frac{A_3}{A_1 + A_2 + A_3} \tag{6}$$

3.4 Deep Ensemble Learning Model

Ensemble learning techniques are proposed to get the final fusion predictions [41,42]. For this purpose, three state-of-the-art classifiers, i.e., CNN, LSTM, and GRU, are used. Fig. 7 depicts the suggested method's deep ensemble module's detailed operation.

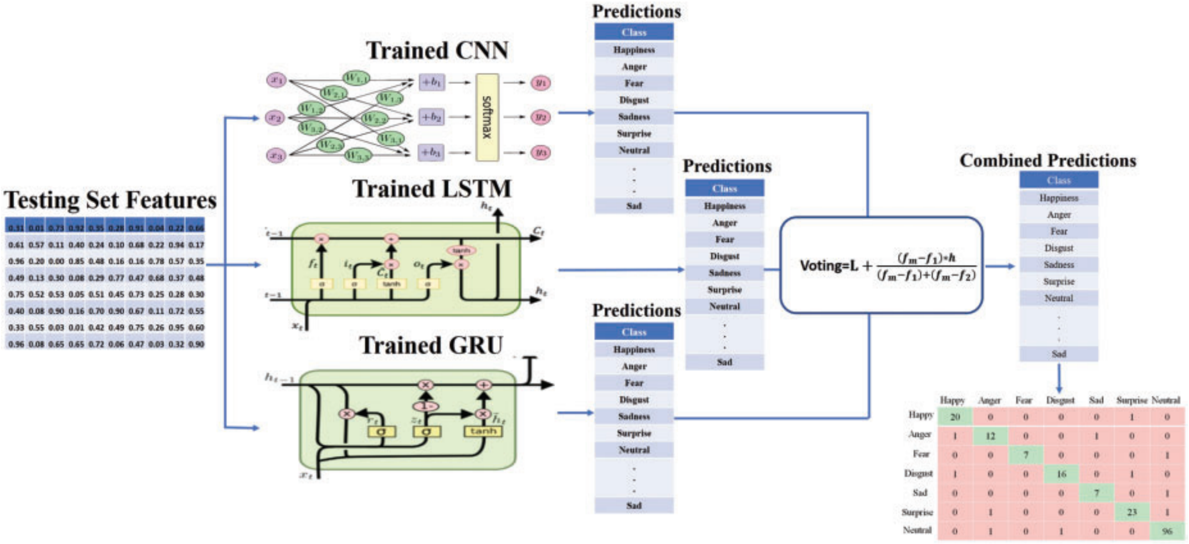


Figure 7: Ensemble learning module

3.4.1 CNN with Softmax Classifier

The CNN model has four layers where; the first layer is features input layers, which are connected to a fully connected layer; similarly, the Softmax layer relates a fully connected and classification output layer [43,44]. The architecture used in this paper can be shown in Fig. 8.

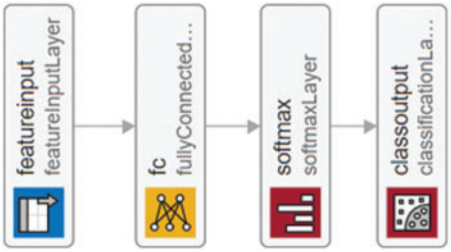


Figure 8: CNN with softmax classifier

3.4.2 LSTM

The second model in the deep ensemble learning module is the LSTM classifier with layers {‘Sequence Input’, ‘LSTM Layer’, ‘Fully Connected’, ‘Softmax’, ‘Classification’} as shown in Fig. 9.

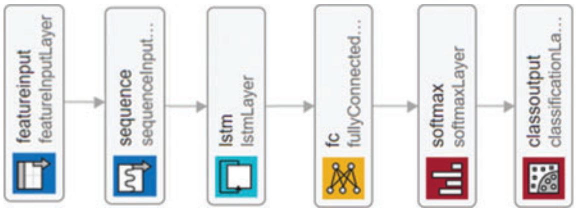


Figure 9: LSTM classifier layers

3.4.3 GRU

In contrast, the last model in the module is the GRU classifier in which the layer arrangement is as shown in Fig. 10, {‘Sequence Input’, ‘GRU Layer’, ‘Fully Connected’, ‘Softmax’, ‘Classification’}.

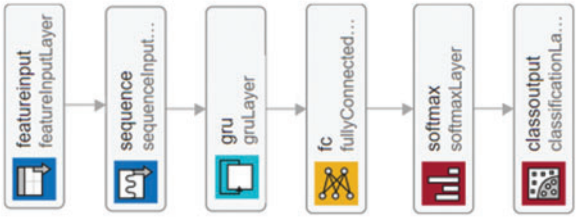


Figure 10: GRU classifier layers

3.4.4 Voting

A majority voting scheme is applied to the output of three classifiers. We select the best flavor (plurality voting) out of three from the majority voting. In this flavor, the output with maximum is selected as the final prediction, which is emotion recognition in our proposed model. A detailed and reasonable examination of the majority voting procedure is employed in the proposed methodology. Mathematically plurality voting can be defined by the following equation,

$$\sum_{f=1}^{F} d_{f,c*} = max_c \sum_{f=1}^{F} d_{f,c} \tag{7}$$

where F is the number of classifiers and C represents the number of classes. Here in our case, f = 1,2, and 3 are CNN Softmax classifier, LSTM, and GRU, respectively, while C = 0,1 . . . 6 represents seven classes in each classifier, i.e., Anger, Happiness, Fear, Disgust, Surprise, Sadness, and Neutral.

4 Experiments and Results

This part of the article will go through how the proposed work was implemented, and the results produced employing the proposed method. Furthermore, using state-of-the-art FER methodologies with full comparative study of the suggested approach is carried out, analyzing both quantitative and qualitative assessments. Two benchmark datasets are used, and each database in the proposed system is split into training and testing sets. MATLAB R2021a was utilized for all simulations in the suggested technique which was running on a work station PC with dual xeon CPUs, a 48 GB DDR4 Ram and Invidia 2080Ti 11GB GPU on windows 10 operating system. The detailed explanation of each of the two databases used in the experiments i.e., RAF and AffectNet, are described in the subsequent sections.

4.1 Datasets

We appraised our suggested framework on the two well-known FER datasets, RAF-DB and AffectNet, to verify the proposed facial expression recognition technique.

4.2 Splitting Data

The dataset is split into two parts: the initial phase utilizes the training data set, which has a size of 70%, and the following step uses the test data set, which has 30% (holdout splitting). We have carried out several performance studies using the most widely used deep learning models in the industry.

4.3 Evaluation Metrics

We compared and assessed the performance in automatically identifying facial expressions in terms of efficacy and efficiency.

4.3.1 Effectiveness Metrics

Formulas (8)–(11) represent the Accuracy (A), Precision (P), Recall (R), and F1, respectively, to examine the effectiveness of the suggested approach.

$$A = \frac{TP + TN}{TP + TP + FP + FN} \tag{8}$$

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{11}$$

4.3.2 Efficiency Metrics

We also consider execution time for comparing the presented methodology using three different classifiers. In the coming section, we elaborate on different experiments and the results on said datasets.

4.4 Results

4.4.1 Experiments on RAF Dataset

Experiments on the RAF dataset are performed using holdout (70/30%) splitting. 12,271 (70%) images are used as a training set in the first step, and 3068 (30%) images are used for testing. Three state-of-the-art classifiers, i.e., CNN, LSTM, and GRU, are trained on a 70% trainset consisting of concatenated CNN features and their labels. For performance analysis of the three classifiers, the models are evaluated using the 30% image features data. A detailed quantitative analysis is performed using accuracy, precision, recall, and f-measure, as performance evaluation metrics. An average accuracy achieved by the CNN model on the RAF dataset is 91%, the accuracy achieved by LSTM is 94%, and similarly, the GRU accuracy is 91%. The fourth accuracy is the output of three classifiers based on ensemble learning and a voting scheme. After applying the majority voting on the three predicted labels vector, a new label vector is created, consisting of deep ensemble classification results. The accuracy achieved by the deep ensemble model is 91.66%.

Tab. 1 illustrates the model's accuracy achieved by the proposed model for the RAF dataset. Figs. 11 to 13 represents the confusion matrix for the CNN Softmax, LSTM, and GRU models and Figs. 14 to 16 represent the accuracy/loss graphs for RAF Dataset.

Table 1: Obtained accuracies for RAF dataset

Classifier model	Accuracy
Softmax classifier	91%
LSTM	94%
GRU	91%

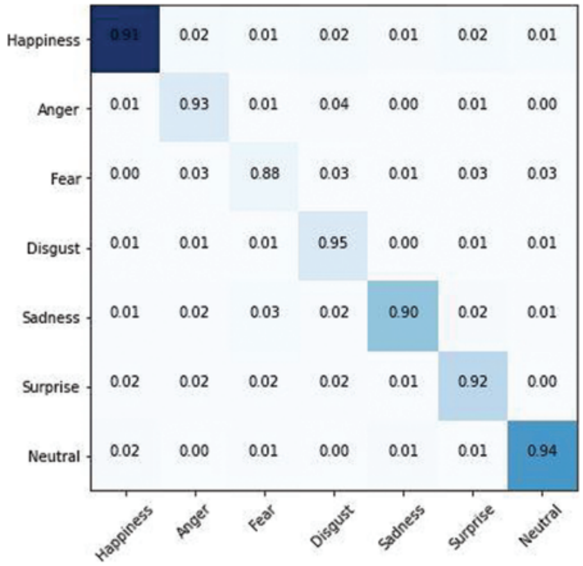


Figure 11: CNN Softmax Classifier

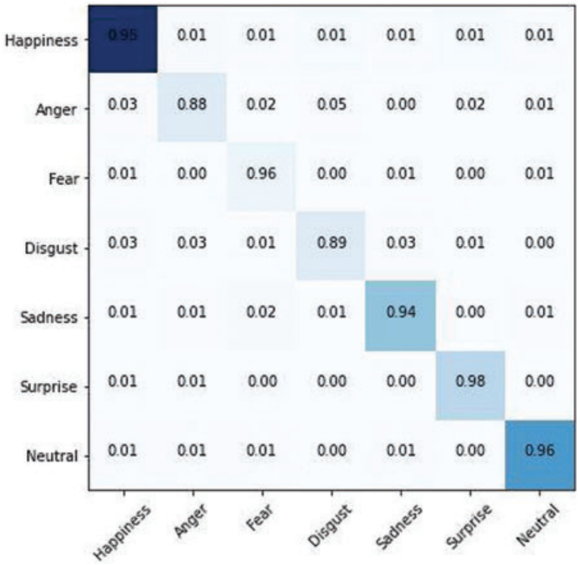


Figure 12: LSTM Classifier

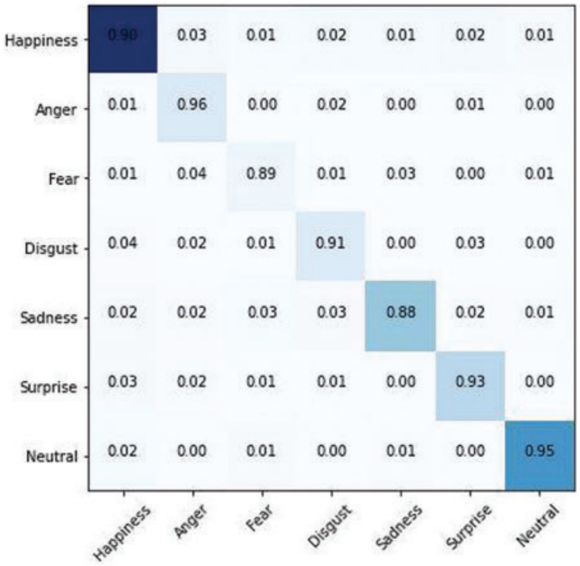


Figure 13: GRU Classifier

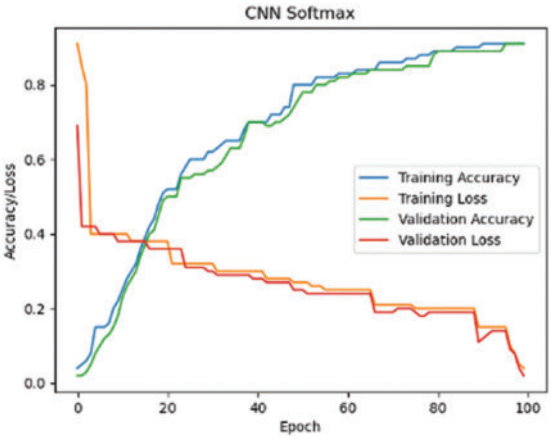


Figure 14: CNN Softmax Classifier

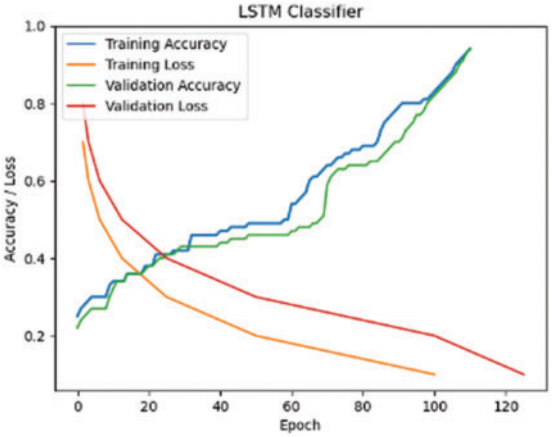


Figure 15: LSTM Classifier

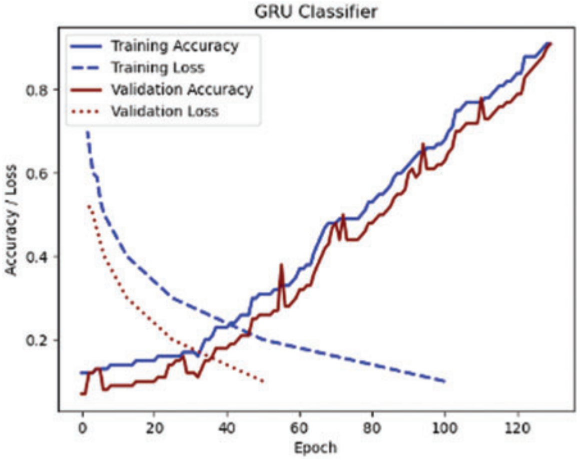


Figure 16: GRU Classifier

Figs. 15–17 demonstrate the training accuracy and loss curves achieved using this method after training 100 epochs for CNN Softmax and 120 epochs for LSTM and GRU. The accuracy curve is observed, and the training and testing accuracy of the model are stable at more than 91 percent after 100 epochs for CNN Softmax, and steady at 94 percent and 91 percent for LSTM and GRU after 120 epochs, respectively, showing that the model has strong classification results.

4.4.2 Ensemble Classification Results for RAF Dataset

After applying the majority voting on the three predicted labels vector, a new label vector is created, consisting of deep ensemble classification results. The confusion matrix obtained for the ensemble classifier is also shown in Fig. 17. The two most perplexing expressions are Disgust and Fear, with Surprise being readily confused with Fear owing to facial similarities, and Disgust being mostly confused with Neutral due to the subtlety of the face. The confusion matrix demonstrates that the suggested method's overall performance is excellent.

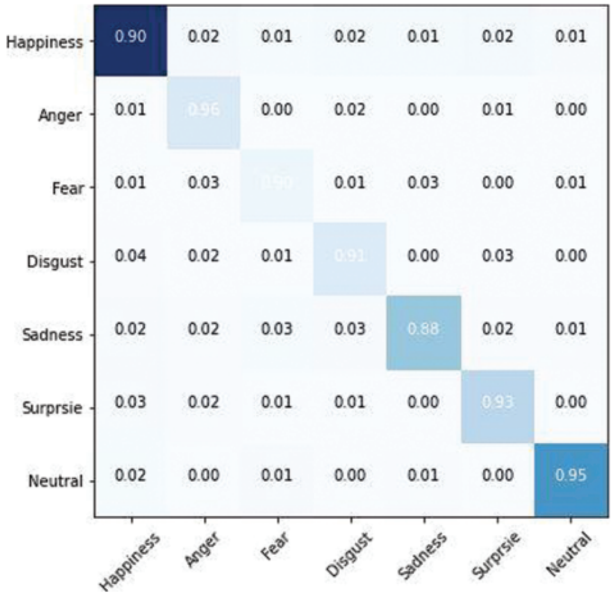


Figure 17: Ensemble confusion matrix for RAF dataset

Tab. 2 provides the effectiveness assessment findings from ensemble classifier models in terms of accuracy, precision, recall, and F-1 score for the basic seven emotions.

Table 2: Accuracy, precision, recall, and f1 score for the RAF dataset

Emotions	Accuracy (A)	Precision (P)	Recall (R)	F1-score
Happiness	94.88%	0.9	0.96	0.93
Anger	98.07%	0.96	0.75	0.84
Fear	98.40%	0.9	0.61	0.73
Disgust	98.27%	0.91	0.79	0.84

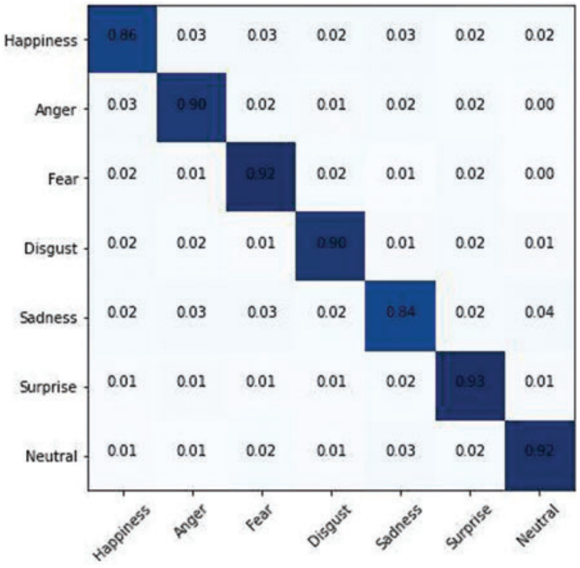
(Continued)

Table 2: Continued

Emotions	Accuracy (A)	Precision (P)	Recall (R)	F1-score
Sadness	97.32%	0.88	0.95	0.91
Surprise	97.91%	0.93	0.88	0.91
Neutral	98.37%	0.95	0.98	0.96

4.4.3 Experiments of AffectNet Dataset

The AffectNet dataset was subjected to the same procedures as the RAF dataset. The only difference is the number of images utilized in the first step is that 291,651 images are used for training, while 4000 images are used for testing. CNN, LSTM, and GRU are three state-of-the-art classifiers that are trained on a 70% trainset of concatenated CNN features and labels. The models are assessed using the 30% image features data for performance analysis of the three classifiers. Various performance assessment indicators, including accuracy, precision, recall, and f-measure, are used to conduct a complete quantitative examination. On the AffectNet dataset, the CNN model has an average accuracy of 89%, LSTM has an accuracy of 65%, and GRU has an accuracy of 61%, as clearly mentioned in Tab. 3. The output of three classifiers based on ensemble learning and a voting mechanism yields the fourth accuracy. Following the application of majority voting to the three predicted labels vectors, a new label vector including deep ensemble classification results is formed. The deep ensemble model achieves a level of accuracy of 71.6%. Figs. 18 to 20 represents the confusion matrix for the CNN Softmax, LSTM, and GRU models for AffectNet Dataset.



	Happiness	Anger	Fear	Disgust	Sadness	Surprise	Neutral
Happiness	0.86	0.03	0.03	0.02	0.03	0.02	0.02
Anger	0.03	0.90	0.02	0.01	0.02	0.02	0.00
Fear	0.02	0.01	0.92	0.02	0.01	0.02	0.00
Disgust	0.02	0.02	0.01	0.90	0.01	0.02	0.01
Sadness	0.02	0.03	0.03	0.02	0.84	0.02	0.04
Surprise	0.01	0.01	0.01	0.01	0.02	0.93	0.01
Neutral	0.01	0.01	0.02	0.01	0.03	0.02	0.92

Figure 18: CNN softmax classifier

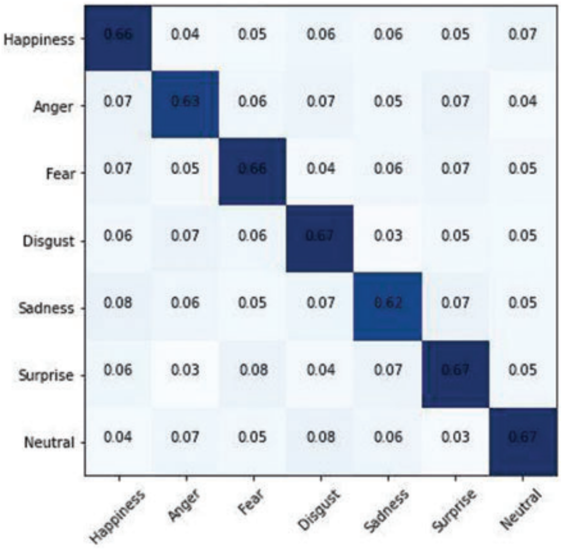


Figure 19: LSTM classifier

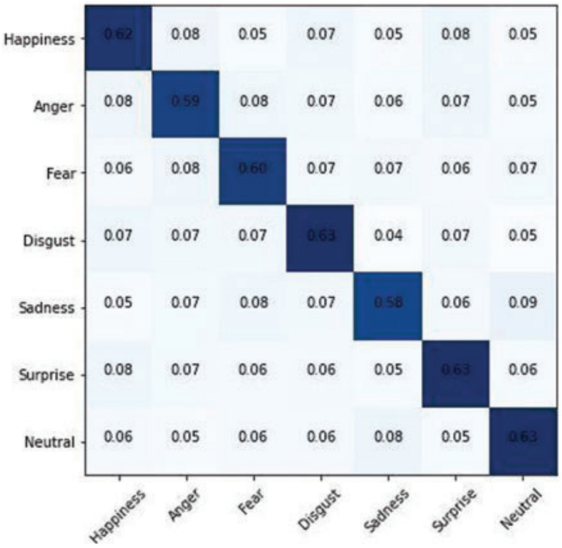


Figure 20: GRU classifier

Table 3: Obtained accuracies for AffectNet dataset

Classifier model	Accuracy
Softmax classifier	89%
LSTM	65%
GRU	61%

Fig. 21 shows the plot for training and validation accuracy/loss phase for the CNN softmax model. The accuracy obtained is 89%. Similarly, training and validation graph for LSTM classifier are shown in Fig. 22. It is proven in Fig. 23 that the loss values for GRU Classifier converge to 0.0047 and 0.0299, in the training and validation loss curves respectively. Furthermore, the training and validation accuracy scores converge to 60.8 percent and 61 percent, respectively. The loss curve drops quickly at first and then progressively saturates. Accuracy tends to saturate beyond a particular epoch, thus increasing the epoch no longer improves accuracy.

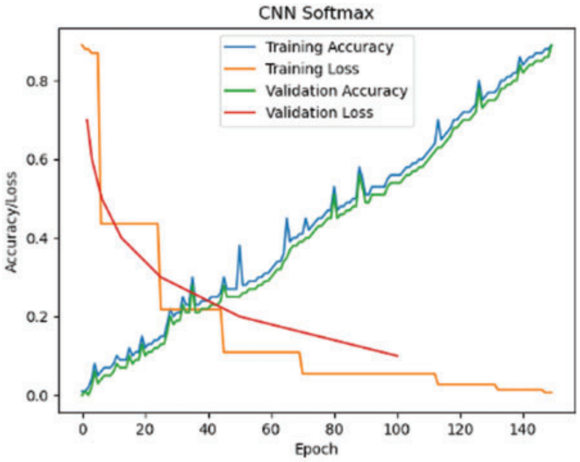


Figure 21: CNN softmax classifier

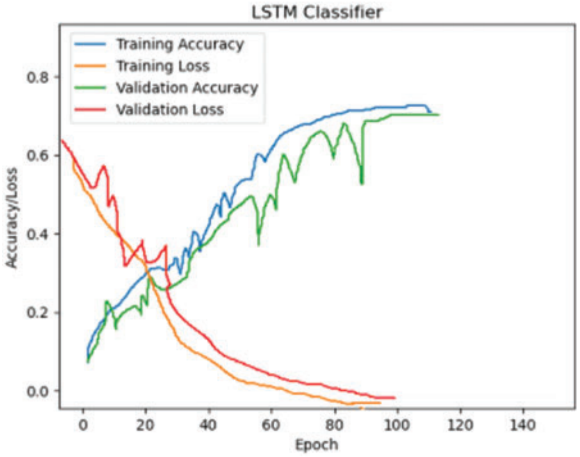


Figure 22: LSTM classifier

4.4.4 Ensemble Classification Results for AffectNet Dataset

Following the application of majority voting to the three predicted labels vectors, a new label vector including deep ensemble classification results is formed. Fig. 24 also depicts the ensemble classifier's confusion matrix of prediction results of seven basic emotions for the AffectNet dataset. It can be seen that our methodology consistently improves the categories of "surprise," "happiness," and "sadness."

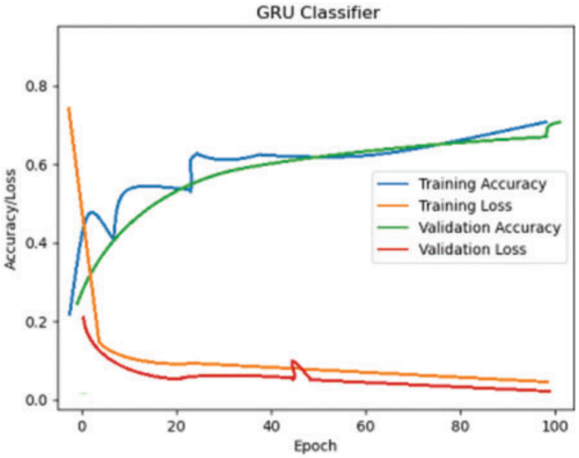


Figure 23: GRU Classifier

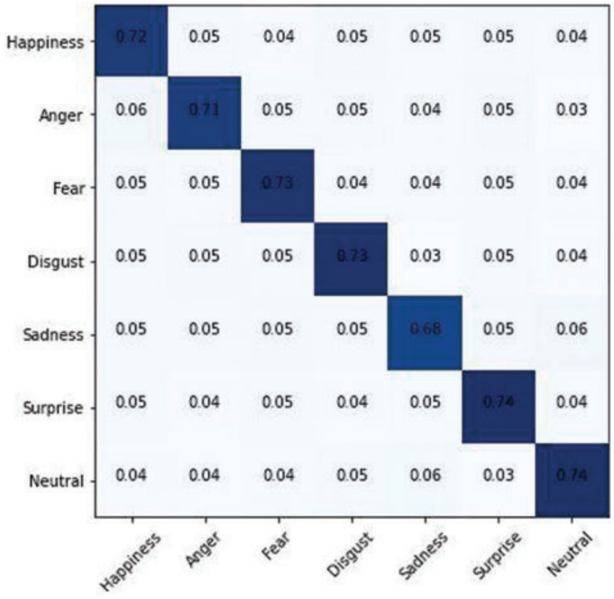


Figure 24: Ensemble confusion matrix for AffectNet dataset

Tab. 4 provides thorough information on the classification performance of ensemble classifier model for individual expression on AffectNet dataset.

Table 4: Accuracy, precision, recall, and f1 score for the AffectNet dataset

Emotions	Accuracy (A)	Precision (P)	Recall (R)	F1-score
Happiness	91.71%	0.72	0.71	0.71
Anger	91.86%	0.71	0.72	0.71

(Continued)

Table 4: Continued

Emotions	Accuracy (A)	Precision (P)	Recall (R)	F1-score
Fear	92.06%	0.73	0.72	0.72
Disgust	92.14%	0.73	0.72	0.73
Sadness	91.57%	0.68	0.72	0.7
Surprise	92.14%	0.74	0.72	0.73
Neutral	92.57%	0.74	0.74	0.74

4.4.5 Comparison of Accuracy with different Methods

Tabs. 5 and 6 depicts the accuracy achieved by our model, when compared to the most powerful competitive approaches. One or two datasets from RAF and AffectNet, respectively, are used to verify most of the presented procedures. The highest new state-of-the-art results were achieved by the proposed method, which were 91.66% and 72.06% on the test for RAF dataset and validation on AffectNet dataset respectively.

Table 5: Test set accuracy on RAF dataset

Method	Average accuracy
ResiDen [45]	76.54%
ResNet-PL [46]	81.97%
PG-CNN [47]	83.27%
Center Loss [48]	83.68%
DLP-CNN [35]	84.13%
ALT [49]	84.50%
gACNN [50]	85.07%
OADN [51]	87.16%
Proposed model	91.66%

Table 6: Validation set accuracy on AffectNet dataset

Method	Average accuracy
VGG16 [37]	51.11%
GAN-Inpainting [52]	52.97%
DLP-CNN [35]	54.47%
PG-CNN [47]	55.33%
ResNet-PL [46]	56.42%
gACNN [50]	58.78%
OADNN [51]	64.06%
Proposed model	72.06%

5 Conclusion

Aiming at the research problem of human facial emotion recognition, we propose a recognition method based on two wild datasets for facial expression recognition. The varied model structures of RestNet50, VGG-19, and Inception-V3 ensure that feature learning followed by feature fusion is diverse. The complementarity of the three feature extraction models was utilized using Ensemble Learning techniques for final expression classification. Based on the performance analysis of pre-trained deep networks, the proposed deep ensemble model with a combination of CNN Softmax, LSTM, and GRU have produced a better accuracy of 91.66% and 72.06% presented in section 4.4.2 and 4.4.3 for RAF and AffectNet Dataset, respectively. The findings demonstrate that compared to the classic technique for FER, the division of FER into several phases, namely feature extraction by state-of-the-art CNN models, feature fusing, classification with deep ensemble model, and then using the techniques of Majority Voting increases the facial expression recognition effect significantly and improve the accuracy rate even more.

We believe the deeper model and inclusion of vision modality in the proposed scheme will further enhance the efficiency of the proposed model.

Acknowledgement: The researchers would like to thank the Prince Sultan University for funding this project.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Gao, W. Zhao, S. Liu, Z. Liu, C. Yang *et al.,* “Facial emotion recognition in schizophrenia,” *Front. Psychiatry*, vol. 12, pp. 1–10, 2021.

[2] Q. Meng, X. Hu, J. Kang and Y. Wu, “On the effectiveness of facial expression recognition for evaluation of urban sound perception,” *Science of The Total Environment*, vol. 710, no. 10, pp. 1–11, 2020.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias *et al.,* “Emotion recognition in human computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[4] T. Akter, M. H. Ali, M. Khan, M. Satu, M. Uddin *et al.,* “Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage,” *Brain Sciences*, vol. 11, no. 6, pp. 734, 2021.

[5] C. Clavel, I. Vasilescu, L. Devillers, G. Richard and T. Ehrette, “Fear-type emotion recognition for future audio-based surveillance systems,” *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[6] M. Leo, P. Carcagnì, C. Distante, P. Spagnolo, P. L. Mazzeo *et al.,* “Computational assessment of facial expression production in ASD children,” *Sensors*, vol. 18, no. 11, pp. 3993, 2018.

[7] S. Shuo and G. Chunbao, “A new method of 3D facial expression animation,” *Journal of Applied Mathematics*, vol. 2014, pp. 1–6, 2014.

[8] S. T. Saste and S. M. Jagdale, “Emotion recognition from speech using MFCC and DWT for security system,” in *Proc. Int. Conf. of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 1, pp. 701–704, 2017.

[9] P. C. Petrantonakis and L. J. Hadjileontiadis, “Emotion recognition from EEG using higher order crossings,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2010.

[10] A. Mollahosseini, D. Chan and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, pp. 1–10, 2016.

[11] P. Liu, S. Han, Z. Meng and Y. Tong, “Facial expression recognition via a boosted deep belief network,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Columbus*, OH, USA, pp. 1805–1812, 2014.

[12] C.-H. Wu, Z.-J. Chuang and Y.-C. Lin, “Emotion recognition from text using semantic labels and separable mixture models,” *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 2, pp. 165–183, New York, NY, USA, 2006.

[13] K. Han, D. Yu and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Proc. Fifteenth Annual Conf. of the Int. Speech Communication Association*, Singapore, pp. 223–227, 2014.

[14] P. L. C. Courville, A. Goodfellow, I. J. M. Mirza and Y. Bengio, *FER-2013 face database*. Montréal, QC, Canada: Universiti de Montreal, 2013.

[15] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba and J. Budynek, “The Japanese Female Facial Expression (JAFFE) database,” in *Proc. Third Int. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 14–16, 1989.

[16] S. Rehman, S. Tu, M. Waqas, Y. Huang, O. Rehman *et al.,* “Unsupervised pre-trained filter learning approach for efficient convolution neural network,” *Neurocomputing*, vol. 365, no. 1, pp. 171–190, 2019.

[17] P. Khorrami, T. L. Paine and T. S. Huang, “Do deep neural networks learn facial action units when doing expression recognition?” in *Proc. IEEE Int. Conf. on Computer Vision Workshop (ICCVW)*, Santiago, Chile, pp. 19–27, 2015.

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, “End-to-end multimodal emotion recognition using deep neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[19] J. F. Cohn and A. Zlochower, “A computerized analysis of facial expression: Feasibility of automated discrimination,” *American Psychological Society*, vol. 2, no. 6, 1995.

[20] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. European Conf. on Computer Vision*, Switzerland, 8689, pp. 818–833, 2014.

[21] 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.

[22] P. Ekman and W. V. Hager, *Facial action coding system: A technique for the measurement of facial movement*. Hove, UK: Psychologists Press, 1978.

[23] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel *et al.,* “Recognizing facial expression: Machine learning and application to spontaneous behavior,” in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2, pp. 568–573, 2005.

[24] J. Whitehill and C. W. Omlin, “Haar features for FACS AU recognition,” in *Proc. 7th Int. Conf. on Automatic Face and Gesture Recognition (FGR06)*, Southampton, UK, pp. 5–10, 2006.

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

[26] P. V. C. Hough, “Method and means for recognizing complex patterns,” *U.S. Patent*, vol. 3, no. 6, pp. 185–188, 1962.

[27] C. Junkai, Z. Chen, Z. Chi and H. Fu, “Facial expression recognition based on facial components detection and hog features,” in *Proc. Int. Workshops on Electrical and Computer Engineering Subfields*, Istanbul, Turkey, pp. 884–888, 2014.

[28] E. Barsoum, C. Zhang, C. C. Ferrer and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” in *Proc. 18th ACM Int. Conf. on Multimodal Interaction*, Tokyo, Japan, pp. 279–283, 2016.

[29] Z. Han, Z. Meng, A. S. Khan and Y. Tong, “Incremental boosting convolutional neural network for facial action unit recognition,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 109–117, 2016.

[30] Z. Meng, P. Liu, J. Cai, S. Han and Y. Tong, “Identity-aware convolutional neural network for facial expression recognition,” in *Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition*, Washington, DC, USA, pp. 558–565, 2017.

[31] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren and A. Cunha, "FERAtt: Facial expression recognition with attention net," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 837–846, 2019.

[32] K. Wang, X. Peng, J. Yang, S. Lu and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 6897–6906, 2020.

[33] K. W. Peng, X. Yang, D. Meng and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[34] X. R. Zhang, J. Zhou, W. Sun and S. K. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.

[35] S. Li, W. Deng and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2852–2861, 2017.

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

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations (ICLR)*, San Diego, CA, USA, pp. 1–14, 2015.

[38] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

[39] C. Szegedy, V. Vanhoucke, S. Io_e, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2818–2826, 2016.

[40] S. Tu, S. U. Rehman, M. Waqas, Z. Shah, Z. Yang *et al.,* "ModPSO-CNN: An evolutionary convolution neural network with application to visual recognition," *Soft Computing*, vol. 25, no. 3, pp. 2165–2176, 2021.

[41] S. Tu, S. U. Rehman, Z. Shah, J. Ahmad, M. Waqas *et al.,* "Deep learning models for intelligent healthcare: implementation and challenges," in *Int. Conf. on Artificial Intelligence and Security*, Cham, Springer, pp. 214–225, 2021.

[42] S. U. Rehman, S. Tu, Y. Huang and G. Liu, "CSFL: A novel unsupervised convolution neural network approach for visual pattern classification," *AI Communications*, vol. 30, no. 5, pp. 311–324, 2017.

[43] S. U. Rehman, S. Tu, Y. Huang and Z. Yang, "Face recognition: A novel un-supervised convolutional neural network method," in *IEEE Int. Conf. of Online Analysis and Computing Science (ICOACS)*, Chongqing, China, IEEE, pp. 139–144, 2016.

[44] S. Rehman, S. Tu, O. Rehman, Y. Huang, C. Magurawalage *et al.,* "Optimization of CNN through novel training strategy for visual classification problems," *Entropy*, vol. 20, no. 290, pp. 4, 2018.

[45] S. Jyoti, G. Sharma and A. Dhall, "Expression empowered residen network for facial action unit detection," in *Proc. 14th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, pp. 1–8, 2019.

[46] B. Pan, S. Wang and B. Xia, "Occluded facial expression recognition enhanced through privileged information," in *Proc. 27th ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 566–573, 2019.

[47] Y. Li, J. Zeng, S. Shan and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. on Pattern Recognition (ICPR)*, Beijing, China, pp. 2209–2214, 2018.

[48] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. European Conf. on Computer Vision (ECCV)*, Amsterdam, The Netherlands, vol. 9911, pp. 499–515, 2016.

[49] C. Florea, L. Florea, M. Badea, C. Vertan and A. Racoviteanu, "Annealed label transfer for face expression recognition," in *Proc. British Machine Vision Conf. (BMVC)*, Bucharest, Romania, 2019.

[50] Y. Li, J. Zeng, S. Shan and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[51] H. Ding, P. Zhou and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *Proc. IEEE Int. Joint Conf. on Biometrics (IJCB)*, Houston, TX, USA, pp. 1–9, 2020.

[52] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu *et al.,* "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5505–5514, 2018.