BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(6), 2021, Article number: e138819
DOI: 10.24425/bpasts.2021.138819

Multi-model hybrid ensemble weighted adaptive approach with decision level fusion for personalized affect recognition based on visual cues

Nagesh JADHAV * **and Rekha SUGANDHI**

MIT ADT University, Pune, Maharashtra, 412201, India

Abstract. In the domain of affective computing different emotional expressions play an important role. To convey the emotional state of human emotions, facial expressions or visual cues are used as an important and primary cue. The facial expressions convey humans affective state more convincingly than any other cues. With the advancement in the deep learning techniques, the convolutional neural network (CNN) can be used to automatically extract the features from the visual cues; however variable sized and biased datasets are a vital challenge to be dealt with as far as implementation of deep models is concerned. Also, the dataset used for training the model plays a significant role in the retrieved results. In this paper, we have proposed a multi-model hybrid ensemble weighted adaptive approach with decision level fusion for personalized affect recognition based on the visual cues. We have used a CNN and pre-trained ResNet-50 model for the transfer learning. VGGFace model's weights are used to initialize weights of ResNet50 for fine-tuning the model. The proposed system shows significant improvement in test accuracy in affective state recognition compared to the singleton CNN model developed from scratch or transfer learned model. The proposed methodology is validated on The Karolinska Directed Emotional Faces (KDEF) dataset with 77.85% accuracy. The obtained results are promising compared to the existing state of the art methods.

Key words: deep learning; convolution neural network; emotion recognition; transfer learning; late fusion.

1. INTRODUCTION

Personalized affective state detection and analysis is one of the non-trivial areas to address in today's world. With the availability of hardware resources and computing power, it is possible to develop deep neural networks to extract minor level details from the images and videos. Many applications of human emotion detection exist, such as drowsiness detections, mood detection, human affective state recognition, to name a few. Visual cues can be used to recognize the emotion in schizophrenia. The study was to understand the process of facial expressions of emotions in schizophrenia using valence, modalities, and genders [1]. The above stated experiment involves intervention of experimenter to analyse the participants and record the findings. Can we automate this process? This can be an important question which is addressed in this paper. Human emotion recognition falls into three categories as shown in Fig. 1. Detecting the human emotion through facial expression is one of the challenging problems to address due to the unavailability of well-balanced datasets, image quality, poor light conditions, occlusions etc. The verbal expressions and physiological signals such as ECG, EEG etc. are the interesting areas of work due to the unambiguous nature of the data. The process of affect detection from visual cues like facial expressions starts with the pre-processing dataset, followed by feature extraction and classification. In the literature, many researchers have used

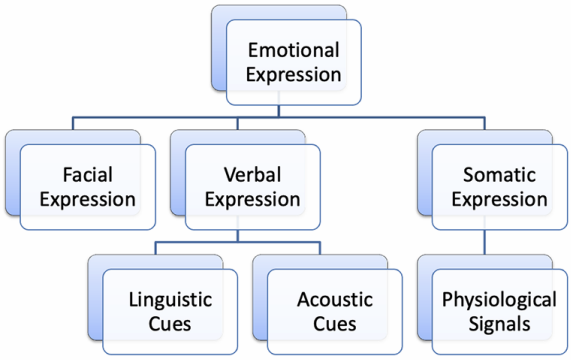

Fig. 1. Emotion categories

different features such as local binary patterns (LPB), histogram oriented gradients (HOG), Gabor filter etc. support vector machine, hidden Markov model, multilayer feed forward perceptron, K- nearest neighbor (KNN), deep neural network (DNN) and convolutional neural network (CNN) are the methods used for classifying human emotions [2].

With the advent of deep learning techniques, CNN is the most preferred approach used for recognizing human emotions. However, developing and training a machine learning model using CNN from scratch is a tedious, time and resource-consuming process. An inductive transfer which is also known as transfer learning can be used to achieve improved results in a short duration of time. Also, in the case of small-sized dataset inductive transfer performs significantly well to achieve state of the art model accuracy. In this paper, we discuss the effect

*e-mail: nagesh10@gmail.com

Manuscript submitted 2021-03-15, revised 2021-07-20, initially accepted for publication 2021-08-21, published in December 2021

of using singleton CNN model and propose an adaptive ensemble approach on multiple models for affective state recognition based on visual cue data perceived from the person [3]. We have also performed various experiments on the different dataset to demonstrate the issues related to model training and parameter fine-tuning. In the deep neural network parameter tuning must be done carefully to achieve better results.

This paper proposes the multi-model hybrid ensemble adaptive weighted approach with decision level fusion for personalized affect recognition based on visual cues using an adaptive ensemble of pre-trained models and CNN, which has shown significant improvement in human affect recognition using facial expressions as primary cue. The further developed model is validated on KDEF dataset.

The remainder of the paper is organized as follows. Section 2 discusses, and reports research and study related to emotion recognition using CNN and transfer learning and describes the basics of the inductive transfer. Section 3 demonstrates the proposed architecture and details. Section 4 discusses the datasets used for experimentation purpose and presents experimental results and summary and finally, Section 5 concludes the paper.

2. RELATED WORK

Many researchers have worked on emotion recognition using CNN considering different datasets. The FER2013 is one of the challenging datasets to work on as it contains the images from the wild. In [4] Mohammadpour *et al.* proposed the CNN model considering facial action units for recognizing human emotion. They used Cohn-Kanade dataset for experimentation and achieved 95.75% accuracy. Cohn-Kanade posed dataset with images captured in controlled laboratory settings. Sang, Dat and Thuan [5] also worked on the FER2013 dataset. Inspired by the VGG model, the authors proposed the four different architectures of 8, 10, 12 and 14 layers to improve the accuracy of the emotion recognition. The presented model focuses on reducing the number of filters, in turn, reducing the number of parameters in the model. Pramerdorfer and Kampel [6] in their state of art discussed different approaches for facial expression recognition using CNN. Authors suggested ensemble for deep convolutional neural networks substantially improve the performance of emotion recognition.

Multi-cue fusion emotion recognition framework is proposed by Yan *et al.* [7] in their research work. The researchers used cascaded CNN and bidirectional recurrent neural network for extracting the dynamic features of the facial emotion. In [8] Rashid suggested the CNN based facial emotion recognition mechanism on JAFFE and Bosphorus dataset. The multiple approaches like decision trees, multilayer perceptron and CNN are used to derive the classification results. The CNN approach resulted in good train and test accuracy. Ruiz-Garcia *et al.* [9] have discussed the CNN approach for recognizing emotion in human faces. They experimented on Karolinska Directed Emotional Faces (KDEF) database. The KDEF database contains 980 images, which are captured in a controlled environment. The proposed approach has two architectures one with reduced deep learning layer and one with split input. Shamim and Ghulam [10] have presented the deep learning approach for emotion recognition from audio-visual emotional big data. Authors suggested the use of CNN for speech signal and video signals. Output two CNNs fused and passed to support vector machine for classification. Vyas *et al.* [11] have surveyed various approaches of facial emotion recognition using CNN and various datasets used for the same. In [12] Zadeh *et al.* proposed a fast facial emotion recognition mechanism using CNN and Gabor filters. The experiments were performed on JAFFE dataset to achieve improved accuracy. Renda *et al.* [13] proposed an ensemble approach for assessing the accuracy of facial expression recognition using the CNN model. Authors also used VGG16 pre-trained architecture in their experimentation. FaceNet2ExpNet, a novel approach, is proposed by Ding *et al.* [14] for the recognition of expression on static images. The model training is done in two phases, the first phase focuses on training convolutional layers while in the second phase fully connected layers are attached to trained convolutional layers. The model was trained and tested on four public datasets, CK+, Oulu-CASIA, TFD and SFEW. Li *et al.* [15] demonstrated the approach for facial expression recognition using transfer learning on small databases. Feature transfer learning approach is used for transferring feature by minimizing feature distribution distance between the source and target datasets. Wang *et al.* [16] have discussed transfer learning with CNN approach for image classification.

The combination of HOG for feature extraction and SVM for pre-classification is used. The pre-classification results are further used for transfer learning using Alexnet [16]. Lee *et al.* [17] developed a system to recognize human emotions based on facial expressions using a webcam. Authors used deep learning-based CNN approach for the training of the model on CK+ database. Person-specific emotion recognition using transfer learning is implemented by Chen *et al.* [18]. Transfer learning is done using a boosting based approach for person-specific modelling. The further transductive approach is used for facial expression recognition. Fan, Lam and Li [19] in their research work have proposed a multi-region ensemble approach using CNN for facial expression recognition. AFEW 7.0 and RAF-DB datasets have been used for the experimentation. The proposed model is based on Alexnet and VGG16 pre-trained models. Residual Network (ResNet) used in the model implementation is a winner of the ImageNet challenge in 2015, is a deep learning model consisting of 150 plus layers. The basic building block of residual learning is shown in Fig. 2 [20]. The researchers in [21] have developed a system to detect fatigue

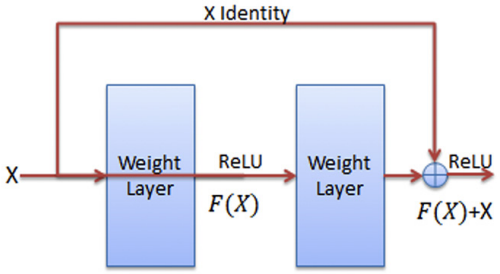

Fig. 2. Residual block structure

symptoms in driver using transfer learning. Pre-trained Alexnet is used to derive desired results. The accuracy achieved is above 90%. Lukasik *et al.* [22] proposed convolutional neural network to detect Latin handwritten characters with diacritics. The overall accuracy achieved is 96%. The convolutional neural has been considered as one of the important breakthroughs in the development of deep learning approaches. CNN is an important building block of deep architecture along with restricted Boltzmann machine (RBM), auto-encoders and recurrent neural networks (RNN). The typical architecture of CNN is given in Fig. 3. It is the architecture of alternating convolutional layer, ReLU and Pooling layers followed by fully connected and output layer [23].

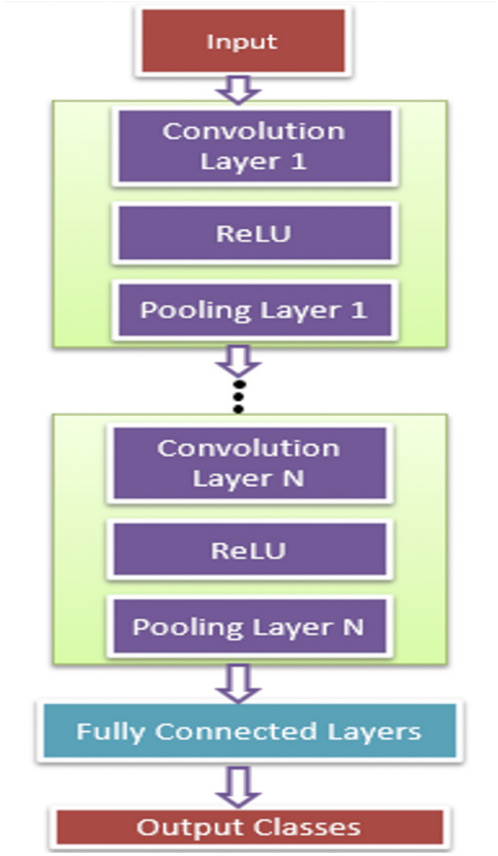

Fig. 3. CNN architecture

Convolutional layer: the convolution is defined as a mathematical operation between the input image X and kernel or filter k to generate convolution output as $Conv_{\text{out}}$, refer Eq. (1) such as,

$$Conv_{\text{out}}(rows, cols) = (X * k)_{rows,\, cols} = \\ = \sum_i \sum_j k(i, j) * X(rows - i, cols - j). \tag{1}$$

The rectified linear unit (ReLU) activation function used to train the network produces the result with reasonable sparsity. ReLU is expressed as,

$$f(z) = \max(0, z). \tag{2}$$

Pooling layer: the main intention of the pooling layer is to reduce the number of parameters i.e., down-sampling, in the model and pace up the calculations. Given a sampling window with (m * m) size, after one down-sampling results into the feature map of size (1/m * 1/m). The pooling layer expression is represented as,

$$x^i = \varnothing\left(\gamma^i * \text{ds}\left(x^{(i-1)}\right) + g^i\right), \tag{3}$$

where, $\varnothing$ is an activation function, γ^i is multiplicative bias, while g^i is additive bias. $x^{(i-1)}$ represents the first feature map of the first layer and ds is a down-sampling function. The pooling layer supports functions namely max pooling and average pooling [23]. Fully connected layer (FC layer): The fully connected layer has three variations namely FC flatten layer which takes input from the convolutional block and convert it into a single vector for processing. The first FC layer tries to predict correct classification class and output layer distribute the final probability of each class e.g., using softmax activation function as given in the Eq. (4) below

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_i}} \\ z_i = weight * input^T. \tag{4}$$

2.1. Inductive transfer

Being intelligent species humans can transfer the knowledge between the tasks. The knowledge gained in one task can be applied to perform other tasks. More relevant the task, the more effective will be the results. The inductive transfer is also known as transfer learning. In transfer learning information learned in one task is inductively transferred to the target task to improve the learning of the task. When it comes to image processing, almost every deep neural network demonstrates the same behaviour as far as the initial stages of the processing are concerned. The initial layers learn generic image features like Gabor filters, color blobs etc. These features are not specific to any database or the application, but they are generic features to be extracted and learned during the training phase. The inductive transfer process begins with training the base network source dataset and specified task, followed by transferring the learned features to target dataset [24]. Training the convolutional neural network from scratch is computationally expensive and it is unusual to have a training database of appropriate size. We will be now discussing possible transfer learning approaches based on the new database and its size and likeness with the original database.

Transfer learning approaches/scenarios: For small and similar dataset steps are mentioned in Table 1A. In the case of large datasets, the models have less chance of overfitting, so we can fine-tune the weights of all the layers. The steps to be followed are mentioned in Table 1B. When the target database is different and small compared to the base database, steps in Table 1C need to be followed.

Table 1
Transfer learning scenarios

Scenario no.	A	B	C
	Small and similar dataset	Large and similar dataset	Small and different dataset
i.	Remove fully connected (FC) layers of the pre-trained convolutional neural network.	Remove fully connected (FC) layers of the pre-trained convolutional neural network.	At the beginning of the convolutional network, remove most of the pre-trained layers.
ii.	Add new FC layers with the same number of classes in the target dataset.	Add new FC layers with the same number of classes in the target dataset.	Add new FC layers matching number of classes in the target database
iii.	Freeze the weights in the pre-trained network and randomize weights in new FC layers.	Unfreeze the layers of pre-trained network and initialize the weights.	Freeze all the weights from the pre-trained network and randomize weights of new FC layers.
iv.	Train the network and update the weights in FC layers.	Randomize weights in new FC layers.	Train the network and update weights in FC layers
v.		Train the entire network and update the weights.	

3. PROPOSED METHODOLOGY

The proposed architecture is shown in Fig. 4, which is based on the ensemble of hybrid models adaptively. The architecture consists of the first model which is trained on FER2013 dataset with ResNet50 as pre-trained network.

The test accuracy achieved in the first model is 71.25%, which is better than state of the art models like Deep-Emotion [25] with test accuracy of 70.02% designed to test FER203 dataset. The second CNN model architecture is developed on the JAFFE dataset where the images are resized to (128, 128, 3), and the third model is designed and developed on CK+ dataset. Each model can predict separately with very good accuracy; however, these predictions change with the change in the unseen input or visual data of different ethnicity or demographic area. One of the major observations is that the accuracy of individual model drops when they were validated on cross databases. So, to achieve good prediction accuracy the classification results or decisions are adaptively assembled to conclude final affective state of the person. The accuracies of the models normalized, and respective weight are assigned to each model. The final prediction is concluded based on the adaptive weights assigned to each classification decision. Most reliable decision has been assigned with the highest weight to indicate significant contribution in the decision making. The late fusion of the decisions is done adaptively to retrieve final affective state or emotion of the person.

3.1. Datasets

The proposed system uses facial emotion recognition (FER) 2013 dataset, Japanese female facial expression (JAFFE) dataset, and CK+ dataset for experimentation. FER2013 is a large dataset consisting of 35 887 images. Each image is grayscale and is of size (48, 48, 1) pixels. The dataset focuses on seven basic emotions like anger, disgust, happiness, sadness, fear, neutral and surprise. The distribution and count of the samples are shown in Fig. 5. The JAFFE database contains 213 images containing 7 basic human emotions [26]. The dimension of each image is (256, 256, 1). The Cohn-Kanade plus dataset is an extension of CK dataset with validation of labels and improvement in common performance metric. CK+ contains 593 image sequences addressing seven basic emotions plus 'contempt' as additional emotion. Each image is of (640, 490, 3) dimension. [27]. The sample images from the dataset are shown in Fig. 6.

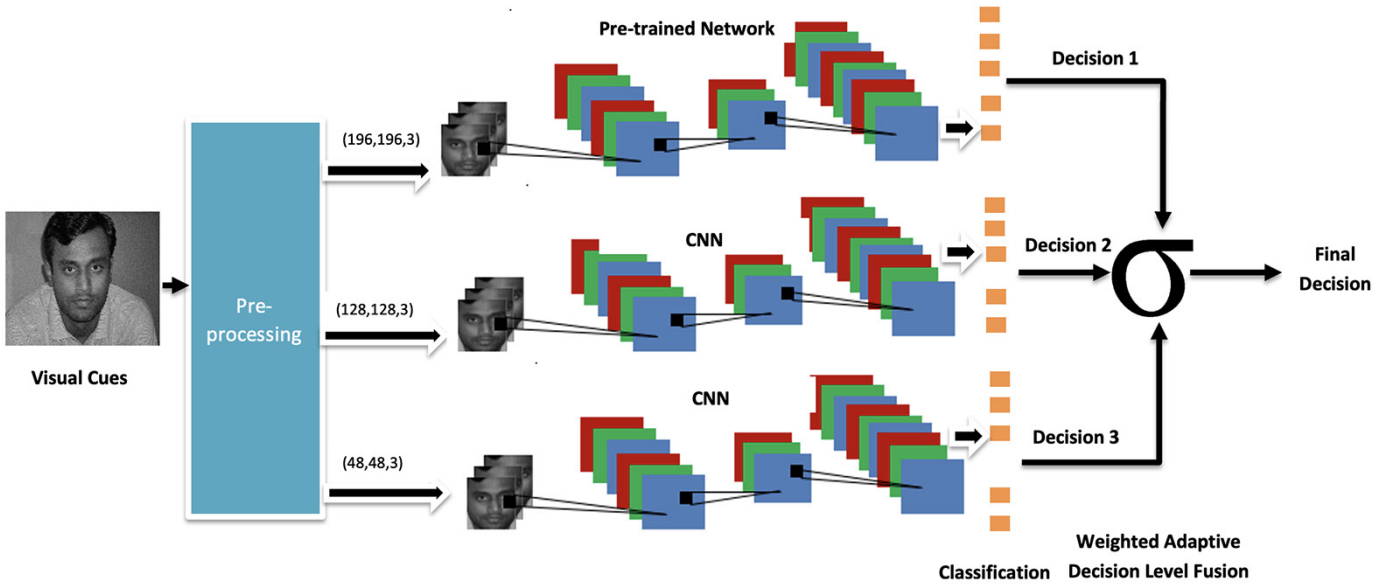

Fig. 4. Proposed System Architecture

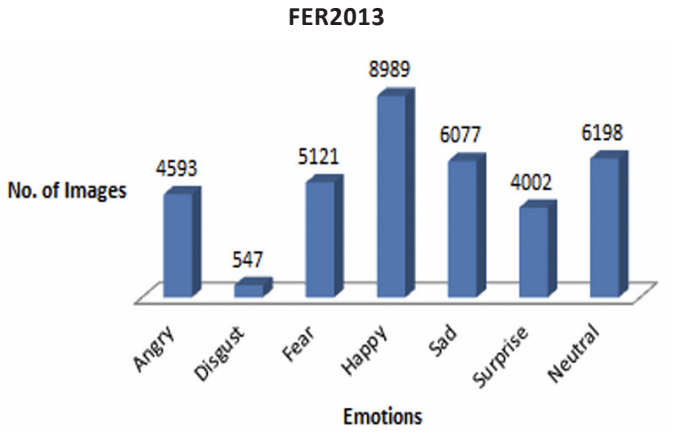

Fig. 5. FER2013 data distribution

Fig. 6. FER2013, CK+ and JAFFE Samples

3.2. Model training with pre-trained network and FER2013 dataset

Recently, transfer learning approach for emotion recognition has been studied and proven to be effective in terms of time for training as well as improved accuracy. The process starts with pre-processing input images, as the default input size and dimension for the ResNet50 is (224, 224, 3) and the FER2013 dataset images are of size (48, 48, 1). The conversion of size and dimension of almost 36k images in FER2013 is labor and resource intensive.

All images were resized to (197, 197, 3). We used Google Colaboratory environment for the pre-processing and training the model. The images were normalized before they were passed to model for the training. The pre-processed images are passed to our model for fine-tuning. By freezing the weights in the initial layers, we tuned the weights in the fully connected layer and passed to dense layer for further processing and classification. The dense layer uses ReLU activation function and softmax at the output layer for classification. The portion of the network is shown in the Fig. 7 below. Due to size constraints, we are showing only initial and final part of model architecture.

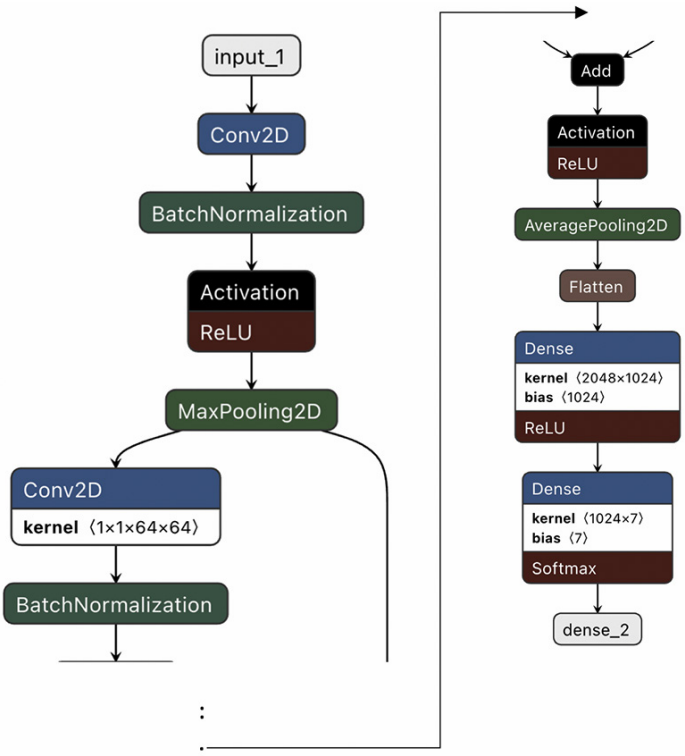

Fig. 7. Model architecture FER 2013 dataset

3.3. Model training with JAFFE dataset

Many images in the FER2013 dataset are not correctly labelled or interpreted wrongly. Also, the disgust emotion has very limited samples to train, so many times model fails to recognize it. The results of cross database testing were not good. To overcome these issues, we trained another deep learning model on JAFFE dataset. We extracted facial landmarks from the images and trained the model accordingly. Each image size in the JAFFE dataset is of size (256, 256, 3). The input images are rescaled to size of (128, 128, 3). The model architecture is shown in Fig. 8.

3.4. Model training with Cohn Kanade plus

To improve on the accuracy, we also worked with CK+ dataset. We used Voila Jones algorithm to extract faces from the images and trained them using convolutional neural network. The model architecture is shown in the Fig. 9.

The dataset images are augmented to increase the dataset size and to prevent the model overfitting. *K*-fold cross-valida-

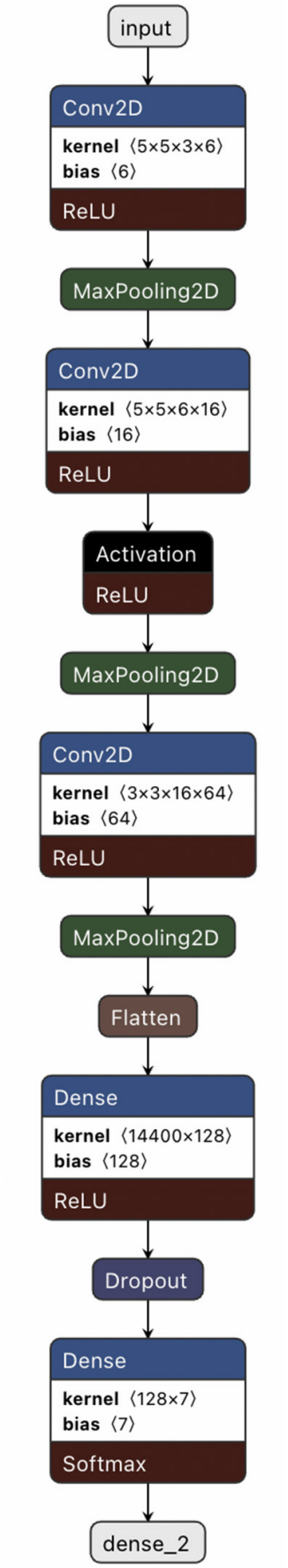

Fig. 8. JAFFE Model

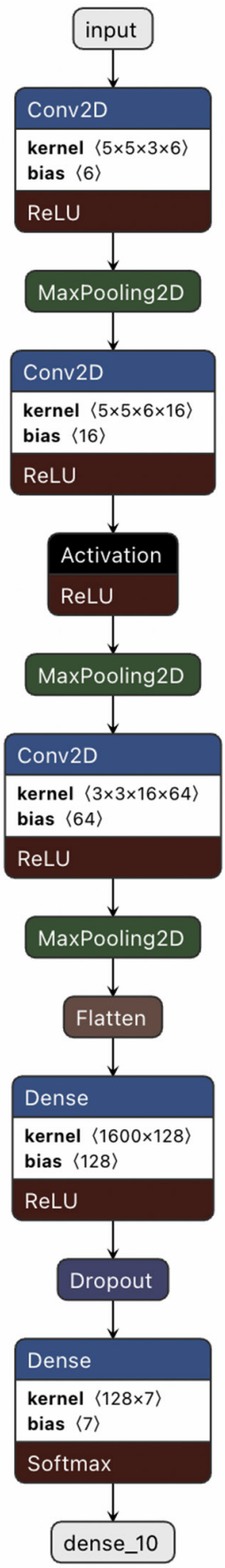

Fig. 9. CK+ Model

tion is used during training to achieve better results. The dataset D is divided into K equal size partitions, $Di, i = 1, 2, 3, \ldots, k$. K-th subpart is a validation set while k–1 parts are used as a training set.

Train1 $= D2 \cup D3 \cup D4 \cup \ldots \cup Dk$ Val1 $= D1$
Train2 $= D1 \cup D3 \cup D4 \cup \ldots \cup Dk$ Val1 $= D2$
Traink $= D2 \cup D3 \cup D4 \cup \ldots \cup D(k\text{–}1)$ Val1 $= Dk$
Er^k = Test error at k-th fold

The estimate of the test error is expressed as,

$$Er = 1/k \sum Er^k. \tag{5}$$

3.5. Data fusion approaches

Data fusion is the process of integrating coherent inputs prior to processing. In the literature, there are three data fusion techniques such as early fusion or feature level fusion, late fusion or decision level fusion and hybrid fusion [28]. Figure 10 depicts

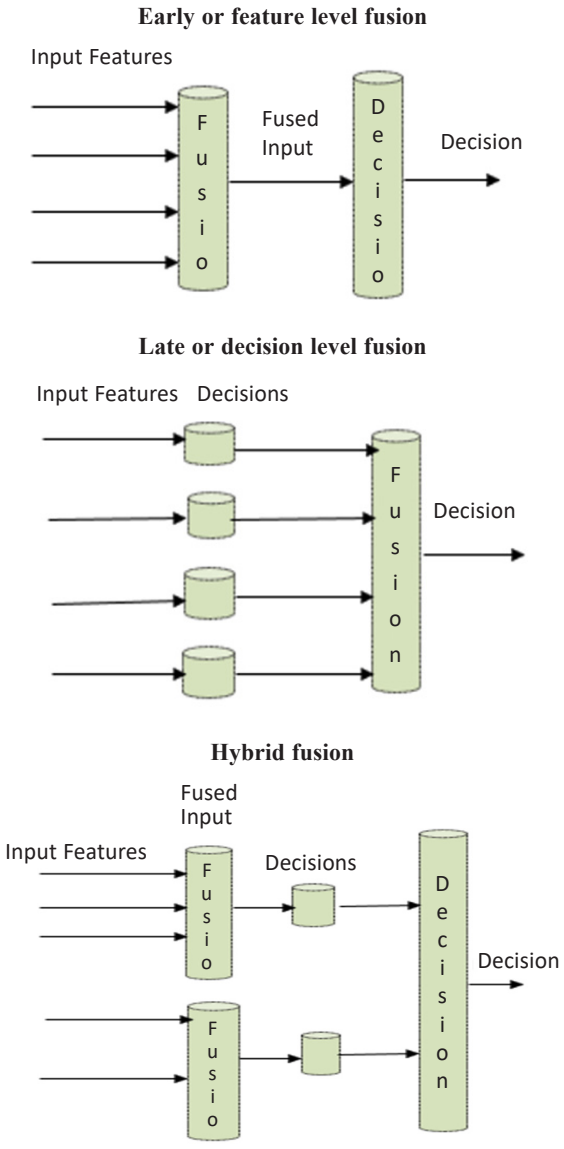

Fig. 10. Data fusion approaches

different data fusion approaches. The feature level fusion happens in the early stage, where the input features of different modalities are combined in vector and passed for further processing. In decision level fusion, decision of each classifier is combined to make a final prediction. Combining early and late fusion approaches lead to the hybrid approach.

3.6. Multi-model ensemble weighted adaptive approach with decision level fusion

The various facial emotion datasets exist in the research domain. Each dataset has its own attributes and emotion classes. The same emotional representation can be perceived differently by the same person [29]. The example is shown in Fig. 11. The disgust emotion is portrayed differently by different people.

Fig. 11. Representation of 'disgust' emotion

Many times, state of the art approaches fail to detect such type of expressions. Also, the ethnicity and facial structure of the person depicting emotion, contributes significantly to emotion recognition accuracy. Hence, relying on a single dataset would result in inappropriate or false predictions. Multi-model ensemble weighted adaptive approach is based on the concept of using a combination of multiple model architectures trained on different datasets. The results from the different architectures are fused adaptively at decision level to make a final decision. Taking a majority vote is always a preferred solution in case of multiple opinions. However, the proposed approach uses adaptive weighted decision instead of votes.

Datasets $= \{T1, T2, \ldots, Tk\}$
Weights $= \{W1, W2, \ldots, Wk\}$
Input feature matrix, $X = [X1, X2, \ldots, Xk]^T$
Class Labels, $Y = \{Y1, Y2, \ldots, Yk\}$
Classifiers $= \{M1, M2, \ldots, Mk\}$
Pre-processing:
Pre-process $T1, T2, \ldots, Tk$ for feature extraction and normalization.

Given input feature matrix and classifiers the decision matrix is represented as,

$$D_{matrix} = \begin{bmatrix} MI, 1(X) & M1, 2(X) & MI, K(X) \\ MJ, 1(X) & MJ, 2(X) & MJ, K(X) \\ ML, 1(X) & ML, 2(X) & ML, K(X) \end{bmatrix}.$$

Each row in D_{matrix} represents, the output of each classifier. The prediction of classifier is the maximum value selected from each row, which is represented as,

$$\text{Pred\_}M1(X) = \text{argmax}\, I_1^k\, D_{\text{matrix}}[M1(X)], \quad (6)$$

$$\text{Pred\_}M2(X) = \text{argmax}\, J_1^k\, D_{\text{matrix}}[M2(X)], \quad (7)$$

$$\vdots$$

$$\text{Pred\_}MK(X) = \text{argmax}\, L_1^k\, D_{\text{matrix}}[MK(X)]. \quad (8)$$

The weights are normalized considering test accuracy as the heuristic,

$$Acc\_Total = \sum_1^k M\_acc(k), \quad (9)$$

where, $M\_acc(k)$ is the test accuracy of each classifier.

$$Wi = \frac{M\_acc(i)}{Acc\_Total} \quad (10)$$

Prediction rules can be written as:

$$Final\_prediction = \text{argmax}\,[(W1 * \text{Pred\_}M1), (W2 * \text{Pred\_}M2), (Wk * \text{Pred\_}Mk)]. \quad (11)$$

To conduct the experiments using multi-modal ensemble weighted adaptive approach, we have input the image to the proposed system architecture. Presented input is pre-processed to extract faces from the input and make it dimension compatible with each model depicted in Figs. 7, 8 and 9. Finally, the result is predicted using the formula mentioned in Eq. (11). The system can predict emotions from still images as well as from the real time videos.

4. EXPERIMENTATION AND RESULTS

The experimental setup includes Google Colaboratory (colab) with 13 GB GPU for training and testing the application. Following are the different scenarios we have considered while performing the experiments on FER2013, JAFFE and CK+ datasets. As a result of all the experiments, we concluded performing inductive transfer for the better results in emotion recognition.

Experiment 1: FER2013 dataset
We have started with stochastic gradient descent (SGD) Nesterov optimizer for the classification of emotions. The experiment settings are given in Table 2.

Table 2
Experiment parameters

Parameters	Values
Train data-size	(29 068, 48, 48, 1)
Validation data-size	(3589, 48, 48, 1)
Test data-size	(3589, 48, 48, 1)
Batch size	64
No. of epochs	100
Loss function	Categorical Cross-Entropy

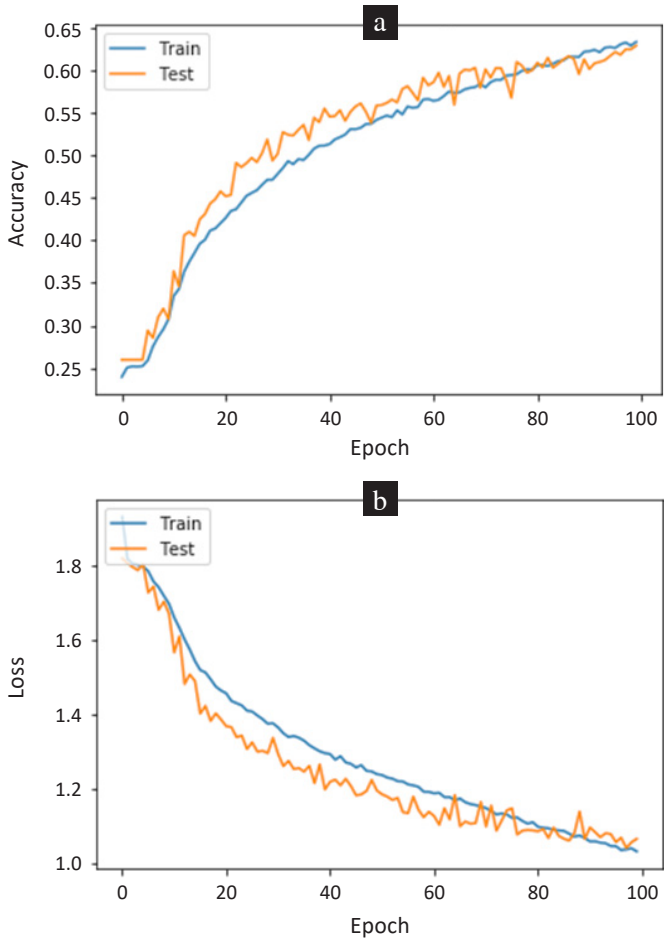

Fig. 12. FER2013_1 Training statistics: a) Accuracy, b) Loss

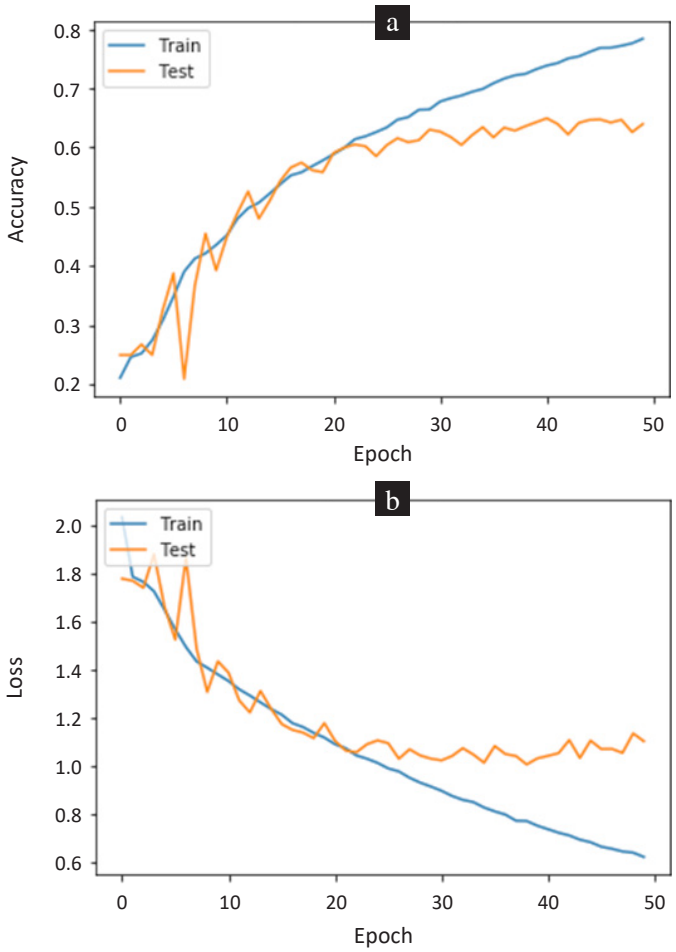

Fig. 13. FER2013_2 Training statistics: a) Accuracy, b) Loss

The time required to run the experiment on FER2013 dataset in colab was 84 minutes and training accuracy achieved was 63.38% and test accuracy was 61.07%. The model accuracy and model loss are shown in Fig. 12.

Experiment 2: FER 2013 dataset
The accuracy achieved in the scenario-1 is not satisfactory. One of the reasons is a non-uniform distribution of the emotion classes in FER2013 dataset. Looking at Fig. 5 we can easily observe that disgust emotion has only 547 samples, which may affect the overall accuracy of the model. Many times, 'disgust' emotion is misclassified as 'anger' or 'fear'. So, we tried testing our model excluding the 'disgust' emotion, assuming we may have increased classification accuracy.

The time taken to train the model on google colab with 13GB GPU was 45 minutes. Train accuracy achieved is 78.39% and test accuracy is 64.50%. This model resulted in the overfitting. The model accuracy and loss are shown in Fig. 13a and Fig. 13b respectively.

Experiment 3: FER 2013 dataset
To improve the test accuracy and the drawbacks of the earlier experiments, we have used ResNet50 pre-trained model with VGGFace weights to train the model on FER2013 dataset. The experimentation settings are given in Table 3. The inductive transfer has significantly improved the test accuracy to 71.25%. The confusion matrix and classification report are shown in Fig.14 and Table 4, respectively.

Table 3
Experiment parameters

Parameters	Values
Train data-size	(28 709, 48, 48, 1)
Validation data-size	(3589, 48, 48, 1)
Test data-size	(3589, 48, 48, 1)
Batch size	32
Optimizer	Adam
Loss function	Categorical Cross-Entropy

Experiment 4: JAFFE dataset
The second CNN model is developed on JAFFE dataset. The model was trained and tested on Google Colaboratory with 71.87% test accuracy. The input images are pre-processed, rescaled and augmented as per the need of experiment. The accuracy graph is shown in Fig. 15.

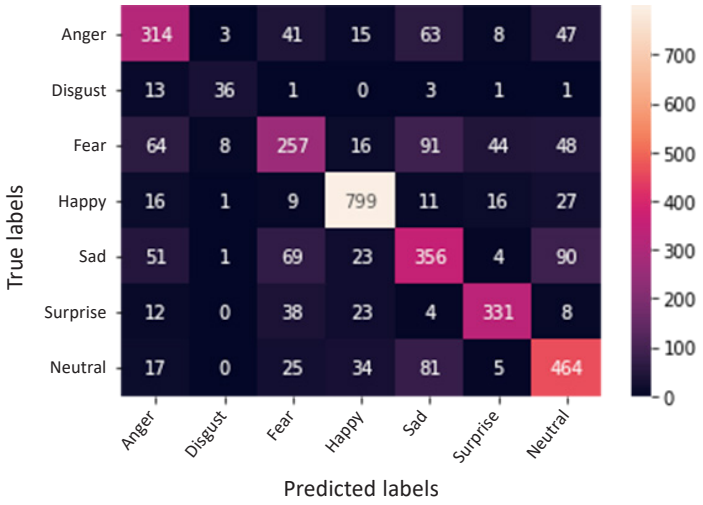

Fig. 14. Confusion matrix

Table 4
Classification report

Emotion	Precision	Recall	F1 Score	Support
Disgust	0.73	0.65	0.69	55
Anger	0.64	0.64	0.64	491
Fear	0.58	0.49	0.53	528
Happiness	0.88	0.91	0.89	879
Surprise	0.81	0.8	0.8	416
Sadness	0.58	0.6	0.59	594
Neutral	0.68	0.74	0.71	626
Accuracy			0.71	3589

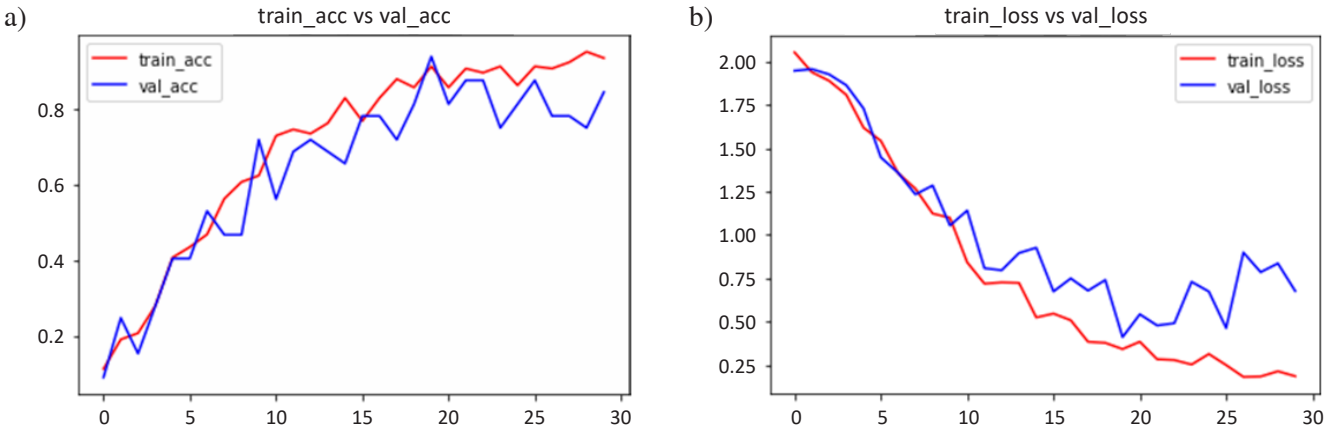

Fig. 15. JAFFE: a) Accuracy, b) Loss

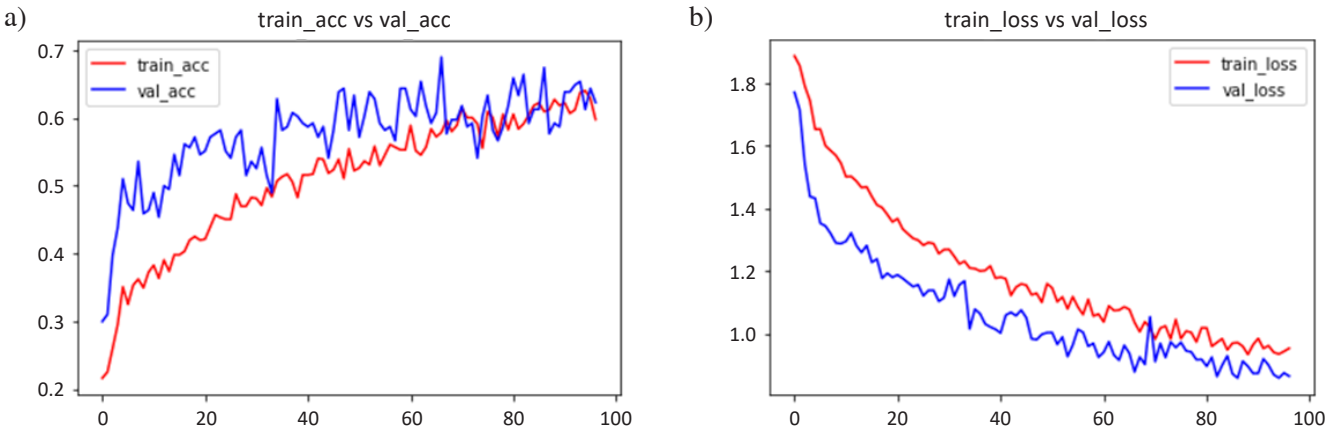

Fig. 16. CK+: a) Accuracy, b) Loss

Experiment 5: CK+ dataset
The CNN model is developed on CK+ dataset. The model is trained using K-fold validation with $k = 5$. The model achieved 84.77% of test accuracy. The accuracy and loss graph are shown in Fig. 16.

4.1. Testing and validation of proposed architecture

The proposed architecture is tested and validated on the Karolinska Directed Emotional Faces (KDEF) dataset [30]. The database contains 490 images depicting 7 basic human emotions. The dimension of each image is $326 * 326$. Each image

is pre-processed for the input size and dimension and passed to the architecture. The prediction accuracy achieved is 77.85%. The sample results are shown in Table 5. The test results on KDEF dataset are displayed in Fig. 17.

Table 5
KDEF Sample Results

Sample	Ground Truth	Prediction
	Anger	Anger
	Sadness	Fear
	Disgust	Anger
	Happiness	Happiness
	Surprise	Surprise

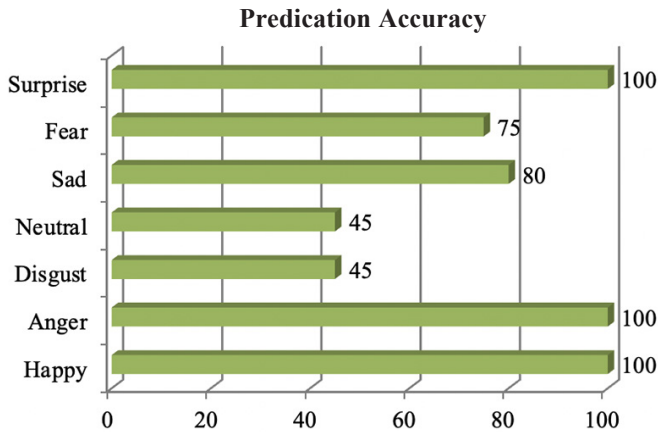

Fig. 17. System prediction accuracy

The overall accuracy on KDEF dataset has been affected due to undertraining on 'disgust' emotion. The trained model has misclassified 'disgust' as 'sadness' for many samples. The same has happened with 'fear', which has been misclassified as 'sadness'. The proposed architecture is further deployed on android mobile for real time usage. Due to the large size of ensemble of models, all the models were quantized to shrink in size. The results can be further improved using fuzzy emotional frameworks and considering other modalities like audio or speech and text to address the issue of ambiguity in affect recognition.

5. CONCLUSIONS

In this paper, we have proposed and implemented a multi-model hybrid ensemble adaptive weighted approach with decision level fusion for personalized affect recognition based on visual cues. The adaptive fusion-based ensemble approach has turned out to be an effective methodology for the small datasets as well as for the imbalanced dataset. The data from different datasets can be exploited to improve the overall accuracy of the system. However, the developed model has limitations in recognizing ambiguous facial expressions. Even for a normal human being it is difficult to differentiate between some of the affects like 'sadness' and 'fear'. Many times, people tend to reserve the portrayal of their affective states intentionally. We have also compared the results of the different CNN models i.e., models for FER2013, CK+ and JAFFE dataset with different hyperparameter configurations. The proposed architecture has achieved 77.85% accuracy on KDEF dataset. The proposed system is deployed on android mobile for real time usage, which can be used in many application areas. The proposed system can be used to identify the depression or anxiety in a person by collecting the visual expressions periodically and analysing the expressions over time. In future, the recognition accuracy can be improved by dealing with ambiguous expressions using fuzzy emotion recognition mechanisms. Also, along with facial expressions, different modalities like speech, physiological signals like EEG and ECG can be fused together to achieve good real time accuracy.

REFERENCES

[1] W. Łosiak and J. Siedlecka, "Recognition of facial expressions of emotions in schizophrenia," *Pol. Psychol. Bull.*, vol. 44, no. 2, pp. 232–238, 2013, doi: 10.2478/ppb-2013-0026.

[2] I.M. Revina and W.R.S. Emmanuel, "A Survey on human face expression recognition techniques," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 6, pp. 619–628, 2021, doi: 10.1016/j.jksuci.2018.09.002.

[3] I.J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015, doi: 10.1016/j.neunet.2014.09.005.

[4] M. Mohammadpour, H. Khaliliardali, S.M.R. Hashemi, and M.M. AlyanNezhadi. "Facial emotion recognition using deep convolutional networks," in *Proc. IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, *Tehran*, 2017, pp. 0017–0021.

[5] D.V. Sang, N. Van Dat, and D.P. Thuan, "Facial expression recognition using deep convolutional neural networks," in *Proc. 9th International Conference on Knowledge and Systems Engineering (KSE)*, *Hue*, 2017, pp. 130–135.

[6] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *ArXiv*, *abs/1612.02903*.

[7] J. Yan *et al.*, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018, doi: 10.1016/j.neucom.2018.03.068.

[8] T.A. Rashid, "Convolutional neural networks based method for improving facial expression recognition," in *Advances in Intelligent Systems and Computing*, *Intelligent Systems Technologies, and Applications 2016. ISTA 2016*, J. C. Rodriguez, S. Mitra, S. Thampi, E. S. El-Alfy (Eds)., vol. 530, 2016, Springer, Cham.

[9] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Deep learning for emotion recognition in faces," in *Artificial Neural Networks and Machine Learning – ICANN 2016*, A.E.P. Villa, P. Masulli, and A.J.P. Rivero (Eds.), vol. 9887, 2016, Switzerland: Springer Verlag, pp. 38–46, doi: 10.1007/978-3-319-44781-0_5.

[10] M. Shamim Hossain and Ghulam Muhammad, "Emotion recognition using deep learning approach from audio-visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019, doi: 10.1016/j.inffus.2018.09.008.

[11] A.S. Vyas, H.B. Prajapati, and V.K. Dabhi, "Survey on face expression recognition using CNN," in *Proc. 5th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2019, pp. 102–106.

[12] M.M. Taghi Zadeh, M. Imani, and B. Majid, "Fast facial emotion recognition using convolutional neural networks and Gabor filters," in *Proc. 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, Tehran, Iran, 2019, pp. 577–581.

[13] A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, "Comparing ensemble strategies for deep learning: An application to facial expression recognition," *Expert Syst. Appl.*, vol. 136, pp. 1–11, 2019, doi: 10.1016/j.eswa.2019.06.025.

[14] H. Ding, S. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, Washington, USA, 2017, pp. 118–126. doi: 10.1109/FG.2017.23.

[15] J. Li *et al*., "Facial Expression Recognition by Transfer Learning for Small Datasets," in *Security with Intelligent Computing and Big-data Services. SICBS 2018. Advances in Intelligent Systems and Computing*, C. N. Yang, S. L. Peng, L. Jain, (Eds.), vol. 895, Springer, Cham, 2018.

[16] Y. Wang, C. Wang, L. Luo, and Z. Zhou, "Image Classification Based on transfer Learning of Convolutional neural network," in *Proc. Chinese Control Conference (CCC)*, Guangzhou, China, 2019, pp. 7506–7510.

[17] I. Lee, H. Jung, C. H. Ahn, J. Seo, J. Kim, and O. Kwon, "Real-time personalized facial expression recognition system based on deep learning," in *Proc. 2016 IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas, USA, 2016, pp. 267–268.

[18] J. Chen, X. Liu, P. Tu, and A. Aragones, "Person-specific expression recognition with transfer learning," in *Proc 19th IEEE International Conference on Image Processing*, Orlando, USA, 2012, pp. 2621–2624.

[19] Y. Fan, J.C.K. Lam, and V.O.K. Li, "Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition", arXiv, 2018, cs. CV, https://arxiv.org/abs/1807.10575v1.

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

[21] J. Chmielińska and J. Jakubowski, "Detection of driver fatigue symptoms using transfer learning," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 66, pp. 869–874, 2018, doi: 10.24425/bpas.2018.125934.

[22] E. Lukasik *et al*., "Recognition of handwritten Latin characters with diacritics using CNN," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 69, no. 1, 2021, article number: e136210, doi: 10.24425/bpasts.2020.136210.

[23] H. Zhang, A. Jolfaei, and M. Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019, doi: 10.1109/ACCESS.2019.2949741.

[24] HackerEarth, "Transfer Learning Introduction Tutorials and Notes: Machine Learning," [Online]. Available: https://www.hackerearth.com/practice/machine-learning/transfer-learning/transfer-learning-intro/tutorial/

[25] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021, doi: 10.3390/s21093046.

[26] M.J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 200–205, doi: 10.1109/AFGR.1998.670949.

[27] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, San Francisco, USA, 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[28] M.F.H. Siddiqui and A.Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," *Multimodal Technol. Interact.*, vol. 4, no. 3, p. 46, 2020, doi: 10.3390/mti4030046.

[29] M.S. Zia, M. Hussain, and M.A.A Jaffar, "Novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier," *Multimed. Tools Appl.*, vol. 77, pp. 25537–25567, 2018.

[30] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska Directed Emotional Faces – KDEF," *CD ROM from Department of Clinical Neuroscience*, Psychology section, Karolinska Institutet, 1998.