Recognizing Human Emotions from Eyes and Surrounding Features: A Deep Learning Approach

Md. Nymur Rahman Shuvo[1]*, Shamima Akter[2]*, Md. Ashiqul Islam[3#]
Shazid Hasan[4], Muhammad Shamsojjaman[5], Tania Khatun[6]
Dept. of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh[1, 3, 4, 5, 6]
Dept. of Bioinformatics and Computational Biology, George Mason University, Fairfax, VA-20110, USA[2]

***Abstract*—The need for an efficient intelligent system to detect human emotions is imperative. In this study, we proposed an automated convolutional neural network-based approach to recognize the human mental state from eyes and their surrounding features. We have applied deep convolutional neural network based Keras applications with the help of transfer learning and fine-tuning. We have worked with six universal emotions (i.e., happiness, disgust, sadness, fear, anger, and surprise) with a dataset containing 588 unique double eye images. In this study, we considered the eyes and their surrounding areas (Upper and lower eyelid, glabella, and brow) to detect the emotional state. The state and movement of the iris and pupil can vary with the various mental states. The common features found within the entire eyes during different mental states can help to capture human expression. The dataset was trained with pre-trained weights and used a confusion matrix to analyze the prediction to achieve better accuracy. The highest accuracy was achieved by DenseNet-201 is 91.78%, whereas VGG-16 and Inception-ResNet-v2 show 90.43% and 89.67%, respectively. This study will provide an insight into the current state of research to obtain better facial recognition.**

***Keywords*—Human emotion recognition; convolutional neural network (CNN); transfer learning; fine-tuning; VGG-16; Inception-ResNet -V2; DenseNet-201**

I. Introduction

Facial expressions are an important nonverbal communication mediums used by humans to express their emotions [1]. In our everyday communication, the facial expression is just positioned next to the tone of the human voice. Different facial expressions are indicators of feelings and it allows a human being to express his/her current emotional state [2]. In Human Computer Interaction, it is crucial to recognize the emotional signs to recognize affective behavior [3], [4]. With the certain forms of thought and affective conducts, physical processes in the human brain changes, subsequently reflects to our eyes and surrounding areas [5], [6]. Therefore, researchers have used signals taken from human brains to understand the emotions, which correspond to the changes through upper and lower eyelid, glabella, and brow of eyes. Therefore, the eyes and surrounding areas can play a vital role in estimating the psychological state of a person. In this article, we work on diagnosing the mental state based on the visible changes in the eyes and its accessory regions to recognize human emotion.

Emotion recognition from facial expressions has incredible importance and wide applications, particularly its functions in human-machine connection frameworks [7]. Numerous studies have been performed to develop and create many facial emotion recognition systems and yet, several common problems still exist in the emotion recognition process. Two major concerns are observed; first, the features are quite sensitive to the changes in noise, illumination, and occlusion. This indicates that a slight change in noise, illumination, and occlusion may reduce the accuracy rate of the recognition process, and second, the large data dimension influences the performance of such systems [8].

The unprecedented development of deep neural networks and convolutional neural networks and the availability of required data have taken the task of classifying and recognition onto another level. Many complex tasks of recognition were thought to be challenging and show less accuracy. With the help of Convolutional Neural Networks (CNN), it becomes possible to achieve higher accuracy [9].

The eye expression could be identified by observing facial tissue signals. However, a few emotions contrast from one another in a couple of discrete facial highlights. This factor additionally relies upon a person's disparities of the subject, for example, degree, frequency, or rate of expression. Based on these, we can say that it is important to develop a facial expressions recognition system which recognizes facial expression in real-time with appropriate accuracy. The purpose of this study is to develop and study the prosperous algorithm of facial expressions recognition and emotion detection on specifically eye images of faces based on deep convolutional neural networks.

II. Related Work

Various methods for the recognition of human emotions from different facial expressions have been developed and analyzed by many researchers. In order to recognize face expressions, researchers applied methods such as Convolutional Neural Network (Deep learning-based algorithm), Viola-Jones algorithm, Haar Cascade Classifier, LBPH, K-Nearest Neighbor. Applying these algorithms, researchers showed various accuracy to predict the outcome and established the improved model which fits for respective dataset(s).

*Both Authors Contributed Equally
#Corresponding Author

Numerous studies have used CNN in their studies to select and optimize active face regions instead of using the whole face area. Researchers [10] have used CNN to extract features from three optimized active face regions i.e. Left eye, right eye, and mouth. Recently, researchers [11] developed a model which was able to predict both primary and secondary emotions by using CNN analysis. Authors utilized fiducial points and a feature selection method to select the relevant features from extracted dynamic features of Neural Network classifier and observed 99% accuracy. Also, researchers [12] presented a system of emotion recognition on video data using both CNN and Recurrent Neural Network (RNN). Study [13] examined emotions and grouped them in six categories by using deep neural network. Authors in [14] showed interest introduce the concept of visual was also based on a CNN structure. Viola-Jones algorithm [15] are using for face detection and deep learning convolutional neural networks for facial expression and emotion recognition. This system has reached a great accuracy rate 92.81%. Overall, the use of CNN seems prominent among researchers to establish great accuracy in the recognition of facial expressions. Beside these studies, some researchers [16],[17] applied deep learning and transfer learning approach to recognize minor visible leaf disease. They proposed a concept of assisted learning where a deep learning model category the emotions from an image into eight categories.

Researchers [18] worked with recognition of seven emotions using dual-feature fusion. They have worked using both texture and geometric features to detect facial expression by using Viola-Jones algorithm [19] in an unconstrained environment and gain average accuracy of 98%. They have used the images from the CMU-MultiPIE database. Researcher [20] used Viola-Jones Haar cascade, Active Shape Model, AdaBoost. They claimed that the systems can provide more accuracy 98% for still images. They worked to achieve better accuracy 97.3% with limited training samples of emotions under varying illumination. For global and local feature extraction researcher [21] used Haar Wavelet Transform (HWT) and Gabor wavelets, respectively. Some researchers have used Local Binary Pattern (LBP) and calculated LBP considering 4-neighbors and diagonal neighbors separately. Their study has shown improvement in the recognition rate on JAFFE, CK, FERG, and FEI face databases in both noisy and noise-free conditions. To analyze seven emotions and calculate the features for a three-dimensional face model, researchers [22] applied k-NN classifier and MLP neural network for feature classification. They gained the highest accuracy from k-NN 95.5%. Their classification accuracy can be affected by real conditions.

Researcher in [23] proposed a multimodal human emotion recognition framework as known as EmotionMeter. In real-life application to improve the chance and durability, they design six electrode placements above the ears to collect EEG signals. Mainly worked on four (happy, sad fear, neutral) emotion using multimodal deep neural networks and achieved best accuracy that is 85.11%. Another study [24] mainly focused on Human Activity Recognition (HAR) and it is a review paper where represents a comprehensive analysis of both handcrafted and learning-based action representations, analysis and discussion on HAR.

Study [25] utilized real-time emotion detection for four basic emotions like happy, sad, anger, fear using five different approaches: AlexNet CNN, Affdex CNN, FER-CNN, SVM, and Multilayer Perception (MLP) and best accuracy achieved from Affdex CNN is 85.05%. Study [26] mainly focused on a novel emotion recognition by using shallowest reliable CNN architecture. They collected data from internet.

Besides, another study [27] mainly focused on multimodal human emotion recognition from eye images and eye movement using two fusion methods: feature level fusion (FLF), BDAE and one classification methods: SVM. For completing this study, they used SEED V datasets and achieved best accuracy from BDAE is 79.63%.

We can categorize the existing methods of emotion recognition from images as dimensional or categorical methods from multilayer hybrid framework [28]. Large scale visual sentiment ontology detections are used to identify adjective-noun pairing [29]. Some researchers recognize emotion based on art feature extraction with art theory [30], [31]. They investigated the shape of features in natural images that influence emotions with visual arts and psychology [32].

III. Proposed System

This is a quantitative applied research based on deep learning approach. After image acquisition, image preprocessing takes part with different parameters for two different dataset splits. Training data goes through various preprocessing techniques (zoom, rotation, flip, and shuffle) to improve data quality by enhancing image features important for the further training part of the system. Resizing (224x224) and rescaling (0-1) techniques are applied to testing data and only the common techniques are applied between training and testing splits. The pre-trained model with customized fully connected layer is trained by training split. The feature extraction part of the model at this time goes frozen and only fully connected layers are trained by the dataset. In fine-tuning, selected convolutional layers that are responsible for extracting features are trained alongside fully connected layers. This training process brings changes in pre-trained weights of the models. Trained models are evaluated by testing data split for comparative analysis. Fig. 1 visualizes the whole process of the system.

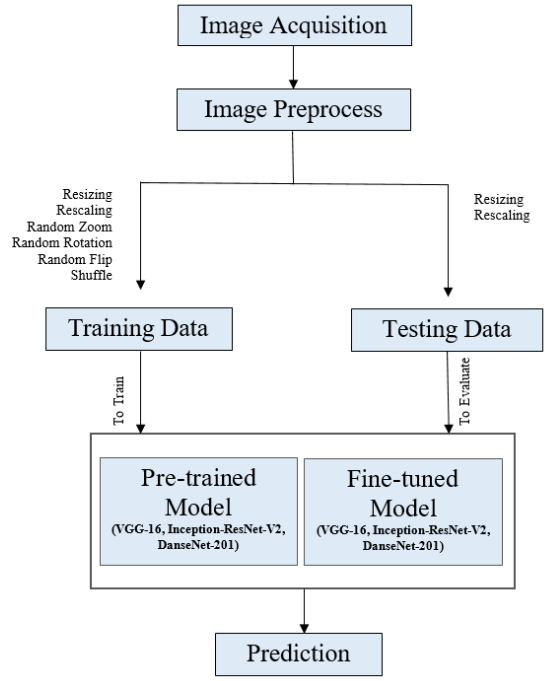

Fig. 1. Proposed Scheme of Emotion Recognition from Eyes and their Surrounding Features.

IV. System Operation

CNN are based on specialized linear operations [2]. It uses convolution instead of matrix multiplication in at least one of their layers. Convolutional layers are used for processing data with a grid-like topology. In computer vision, it is a popular selection for extracting features from visual objects [12].

The convolutional layer takes patches from images known as filters or kernels shown in Fig. 2. These patches are a priority part of an important feature of the entire image. It helps better to understand the features of images than taking the whole image. The filter sizes vary for different layers of network but fix for any individual layer. The dimensionality of the l[st] layer filter represented as Eq. 1.

$$K\,(n) = dim\,(filter) = (fl, fl, nCl - 1) \;\; here, f = filter\ size \tag{1}$$

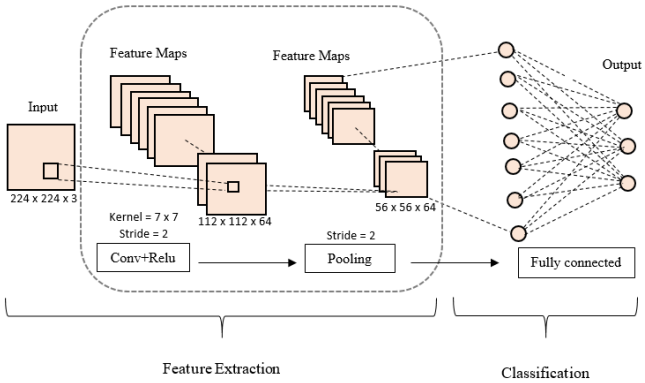

Fig. 2. Feature Extraction and Classification using Neural Network.

The filter shifted over the images with a stride value and computed the feature maps following the mathematical Eq. 2 and Eq. 3.

$$\forall n \in [1,2,3,\ldots nC\,[l]]: conv\,(a\,[l-1], K\,(n))\,x, y = \varphi\,[l]\,(I = 1nW\,[l-1]\,j = 1nH\,[l-1]\,k = 1nC\,[l-1]\,Ki, j, k\,(n)\,ax + i - 1, y + j - 1, k\,[l - 1] + bn\,[l]) \tag{2}$$

$$a[l] = [\varphi\,[l]\,(conv\,(a\,[l-1], K\,(1))), \varphi\,[l]\,(conv\,(a\,[l\,1], K\,(2))), \ldots, \varphi\,[l]\,(conv\,(a\,[l-1], K\,(nC\,[l]\,)))] \tag{3}$$

In Eq. 2 and Eq. 3 a [l-1] denotes input data and the beginning image is a0.nC[l] is the number of filters in the image. Φ and b denotes activation function and bias. a [l] is the output of convolution layer with size of nW[l],nH[l],nC[l]

Rectified linear units (ReLu) activation function is taken part over the feather map that is calculated by convolutional layer. Only the positive values from the feature map remain the same but the negative values are turned to zero in the ReLu function shown in Eq. 4.

$$f(x) = max\,(0, x) \tag{4}$$

The pooling layer aims to pull down sample feature maps generated by following Eq. 3. The feature map value with less impact is neglected and the height value from the feature map is taken for the next sequential layer using the max-pooling function. And calculating the average value from the pooling filter and return is called average pooling. The pooling layer has no parameters to learn.

Eq. 5 is mathematical form of pooling layer,

$$ax, y, z\,[l] = pool\,(a\,[l-1])\,x, y, z = \phi\,[l]\,((ax + I - 1, y + j - 1, k\,[l-1])\,i, j \in [1,2,..,f\,[l]\,]\,2) \tag{5}$$

Here ϕ[l]is the pooling function.

The fully connected layer is situated on top of the model and is used for classification tasks (Fig. 2). Feature vector obtains from the flatten layer and calculates the value to return to the next layer. Eq. 6 shows the node at faithfully connected layer.

$$Zj\,[i] = l = 1ni - 1wj, l\,[i]\,al\,[i-1] + bj\,[i] \tag{6}$$

Here is the weight of the node.

Convolutional blocks:

The convolutional block is a combination of multiple convolutional layers with necessary activation functions and different types of layers. This combination of layers and activation functions works together to extract features from input data. VGG-16 network architecture contains 16 convolutional layers with the same 3x3 kernel size shown in Fig. 3 [33]. The convolutional layers are separated into five blocks. The model increases the number of feature maps as the depth of the network increases. The final layer of each block takes a pooling layer which reduces the size of feature maps.

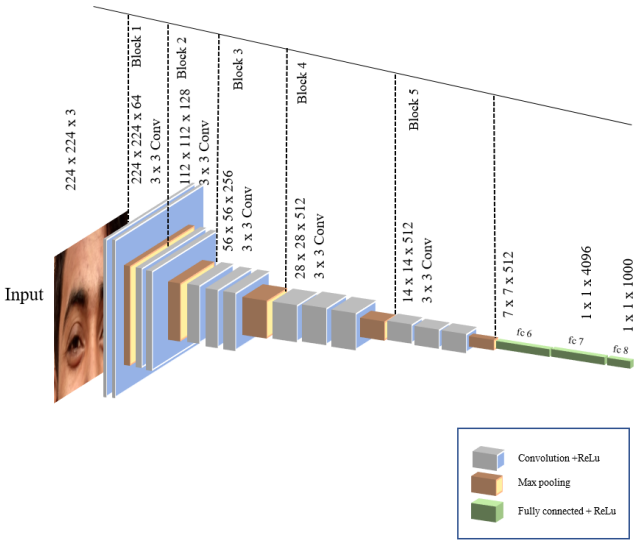

Fig. 3. VGG-16 Network Built-in Blocks.

Inception network breaks the concept of the same filter size in a block [34]. It focuses on different convolutional layers with different filter shape in a single built-in block. Multiple layers subsist in a block with parallel sequences and finally concatenate the output of each sequence of layers. In Fig. 4, the 1x1 convolution layer beginning of a parallel path of a sequence reduces the dimensionality of the input data. Concatenate the output from different filter size sequences facilitated with multi-level feature extraction.

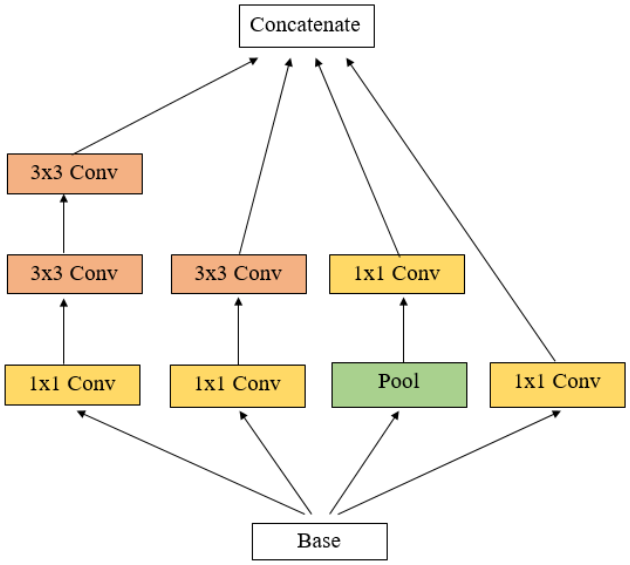

Fig. 4. Schema of INCEPTION BLOCK. Parallel LAYER Sequence with Different Filter Sizes.

Res-Net network fed the output of a convolutional layer to not only the next layer but also ahead of the layer. Then performed element-wise addition. This approach improves the vanishing gradient problem of the network [35]. The Inception-Res-Net network combines the concept of Inception and Res-Net [36]. Fig. 5 shows Inception-Res-Net blocks. This combination can extract multi-level features from images with less vanishing gradient problems. Inception-Res-Net blocks followed by a 1x1 convolutional layer (without activation) scaling up the dimensionality of feature maps before concatenation.

In the Dense block [9], every convolutional layer obtains direct input from every previous layer shown in Fig. 6. Input feature maps from the previous block, at first go through the batch normalization layer which standardizes the input data. After each convolution, the number of channels remains the same and the number of channels indicates the growth rate of a block. After convolution, mapped output feature is sent not only to next convolutional layer but to the rest of the layers of the block.

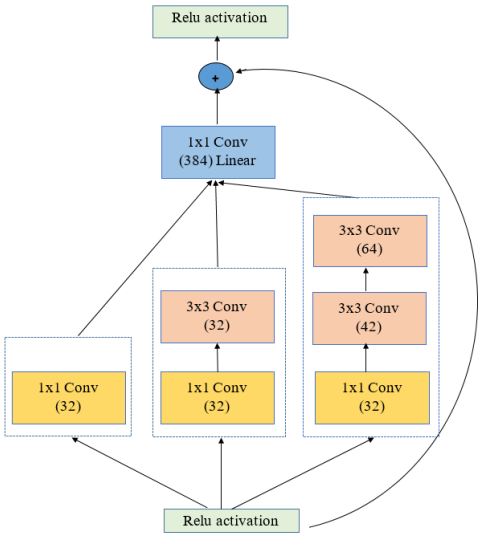

Fig. 5. Schema of Inception-ResNet-V2 Built-in Blocks.

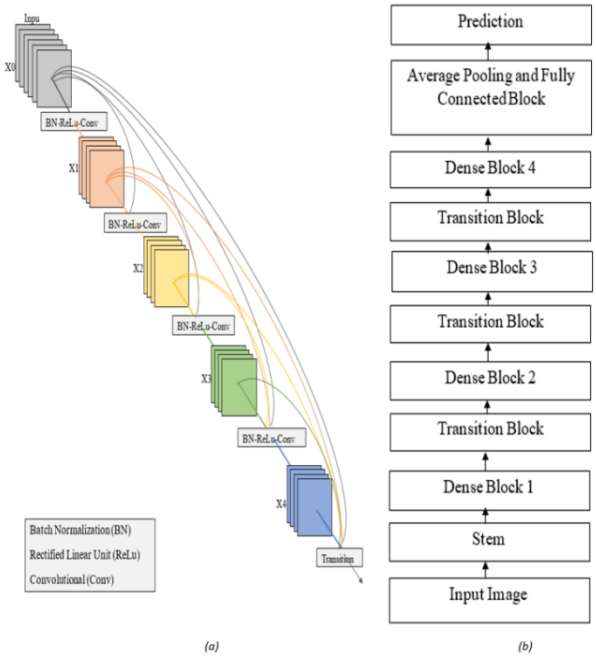

Fig. 6. 5-layer Dense Block Consists of a Growth Rate of 4(a) [6]. Schema of DenseNet-201 Network (b).

A. Transfer Learning

Transfer learning is a machine learning approach that overcomes the isolated learning paradigm of one model. It opens the door of sharing knowledge to solve related problems. Keras applications (VGG-16, Inception-ResNet-V2, DenseNet-201, etc.) are trained with ImageNet dataset of millions of images of thousands of different classes [5]. The trained Keras model weights are transferable to make it easier to solve the related problem. The ImageNet dataset contains 'person', 'individual', 'someone', 'somebody', and more humanistic classes of photos. Transferring this knowledge gained from the ImageNet dataset can boost up the learning performance of selected models with the eye dataset.

B. Fine Tuning

A pre-trained model contains weights in its node for solving any particular problem. The Keras model is pre-trained model trained with thousands of categorical images of the ImageNet dataset. Fine-tuning, a pre-trained model creates opportunities to solve correlated problems by changes in the weight values of the model. Different blocks of convolutional layers of deep neural networks contain different feature pattern recognition abilities. The last layers and blocks contain the most specific feature pattern of objects. The starting convolutional layers and blocks contain general features of objects like edges and shapes. Updating the weights of the upper blocks with the unique dataset can utilize the model more efficiently.

V. Facial Expression Types and Dataset Descriptions

Human countenances are ostensibly the foremost things we see. We rush to differentiate them in any scene, which they command our consideration. Countenance plays a very important role in our daily lives and we express our emotions through this. Plenty of times we do not say anything but just our facial expressions can explain our situation. Although there are 21 or 30 kinds of facial expressions overall, 7 or 8 kinds of facial expressions are considered universal expressions like anger, disgust, fear, happiness, sadness, and surprise, contempt [37]. Once we convey any quiet expression on our face, all the part of the face like nose, eyes, lips, etc. carry a kind of change. It varies in several expressions. We will read human emotions by watching the expression of a specific organ of the whole face.

In this paper, we have focused on 6 universal facial expressions such as anger, disgust, fear, happiness, sadness, and surprise of human emotions by watching eye expressions shown in Fig. 7. The source of the dataset is attached in Kaggle (link: https://www.kaggle.com/mdnymurrahmanshuvo/eye-emotion-dataset-diu) the image mathematically generally represented and showed in Eq. 7.

$$dim\ (image)\ = (nW, nH, nC) \tag{7}$$

Here, nW, nH, nC respectively represent the size of the width, size of height, and several channels of an image.

The dataset quantitative properties are described in Table I(a), (b). The dataset is separated into training and testing parts. In the training dataset, each data class shares an equal amount of data.

Fig. 7. Eye and its Surroundings Feature Condition in Different Emotional States.

TABLE I. (A) EYE-EMOTION DATASET DISCUSSION

Dataset features	Parameters
Total instance	588
Total training data	450
Total testing data	138
Number of classes	6

(B) EYE-EMOTION DATASET DISCUSSION

Name of classes	Number of test data in class	Number of test data in class
Angry	75	27
Disgust	75	15
Fear	75	23
Happy	75	34
Sad	75	21
Surprise	75	18

VI. Model Descriptions

Keras deep learning models are used for prediction, feature extraction and fine-tuning. The models are available with pre-trained weights. Table II contains notable information on some Keras applications used in this study for feature extraction. Depth of deep learning model defines the number of layers exist in the network. The weights of an artificial neural network refer as parameters of it. VGG-16 is a convolutional neural network architecture and was visualized (Fig. 3) with a schematic representation of the VGG-16 network architecture [33]. It improves the performance of Alex-Net by reducing the filter size and increasing the number of channels as the depth of the network. Inception-ResNet-V2 network architecture combines the concept of multi-feature extraction with the reduction of vanishing gradient issues [12]. Fig. 5 concerting the built-in blocks of the network where multiple filter size layers take part (1x1, 3x3) for feature extraction and concatenate the results obtained from parallel layer sequences and the input data from the previous layer (conceptualize from Res-Net architecture). Three Inception-ResNet blocks with a different number and layer combinations take place in the sequential model. The denseness-201 deep convolutional network contains four dense built-in blocks. Three transition blocks followed by the first three dense blocks. A fully connected layer block situated at top of the network shown in Fig. 6.

TABLE II. SELECTED PRE-TRAINED MODEL DESCRIPTION

Model Name	Depth	Number of Built-in Blocks	Parameters	Top-5 Accuracy
VGG-16	23	5	143,667,240	0.901
Inception-ResNet-V2	572	3	55,873,736	0.953
DenseNet-201	201	4	20,242,984	0.923

VII. EXPERIMENTAL ANALYSIS

A. Method and Result Analysis

This study has performed with three different deep convolutional neural network architectures with a dataset of 588 instances. To accelerate the learning process, this study used pre-trained weights to the network. Use of Adam's optimization function with a learning rate of 0.0009 and another parameter like beta_1-2, epsilon remains the default. 450 epochs get better performance than the nearest numbers of it, where stepper epoch is taken as 10. For training and testing performance measurement, a confusion matrix is used to generate the accuracy ratio. Eq. (8) formula used to calculate the accuracy.

$$Accuracy = TP + TN/TP + TN + FP + FN \quad (8)$$

After the execution of the models with the same dataset, this study found different training and testing accuracy with different methods shown in Tables III and IV. Training accuracy refers the accuracy of the model with training data and testing accuracy refers the accuracy obtain from the model while applying testing data. By freezing the trainable layers, the DenseNet-201 network achieves the height accuracy of 91.78% and 89.13% training and testing terms, respectively.

The noise of a dataset is also trained by a model which creates a performance gap between training and testing accuracy. This term is called overfitting in the field of machine learning. The DenseNet-201 model contains less overfitting of 0.0265 compared to other models.

Fig. 8 contains a training and testing accuracy graph of three different models where the DenseNet-201 graph is smoother than other network architecture. From Fig. 10(b) we observed the Inception-ResNet-V2 loss curve and indicate more overfitting than other two models.

After evaluating the model with test data, the model predicts the class name based on its learning. The predicted result could be True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FS).

TABLE III. RESULT OBTAINED FROM THE PRE-TRAINED

Models	Training Accuracy	Testing Accuracy	Overfitting
VGG-16	0.9043	0.8768	0.0275
Inception-ResNet-V2	0.8967	0.8551	0.0417
DenseNet-201	0.9178	0.8913	0.0265

TABLE IV. RESULT OBTAINED FROM FINE-TUNED MODELS

Model Name	Training Accuracy	Testing Accuracy	Overfitting
VGG-16	0.8432	0.8106	0.0326
Inception-ResNet-V2	0.8401	0.7806	0.0595
DenseNet-201	0.8621	0.8170	0.0451

Precision refers to the ratio of total correctly predicted positive values (TP) to total predicted positive values (TP+FP).

$$Precision = TP/TP + FP \quad (9)$$

Recall refers the proportion of total correctly predicted positive values (TP) to total actual positive values in dataset (TF+FN).

$$Recall = TP/TP + FN \quad (10)$$

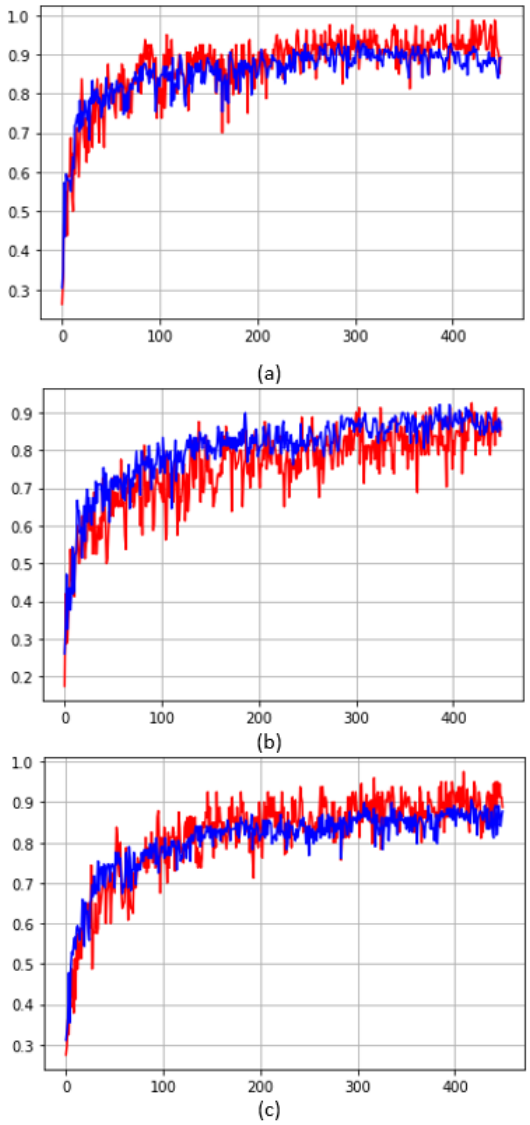

Fig. 8. Accuracy Graph of DenseNet-201 (a), InceptionResNet-V2 (b) and VGG-16(c).

TABLE V. PREDICTION RESULT ANALYSIS OF DENSENET-201 MODEL

Class Name	Precision	Recall	F1-score
Anger	0.89	0.89	0.89
Disgust	0.77	0.67	0.71
Fear	0.96	1.00	0.98
Happy	0.91	0.85	0.88
Sad	0.90	0.90	0.90
Surprise	0.86	1.00	0.92

TABLE VI. PREDICTION RESULT ANALYSIS OF INCEPTION-RESNET-V2 MODEL

Class Name	Precision	Recall	F1-score
Anger	0.83	0.93	0.88
Disgust	0.75	0.80	0.77
Fear	0.79	0.96	0.86
Happy	1.00	0.85	0.92
Sad	0.89	0.81	0.85
Surprise	0.81	0.72	0.76

TABLE VII. PREDICTION RESULT ANALYSIS OF VGG-16 MODEL

Class Name	Precision	Recall	F1-score
Anger	0.89	0.93	0.91
Disgust	0.76	0.87	0.81
Fear	0.92	0.96	0.94
Happy	0.96	0.79	0.87
Sad	0.81	0.81	0.81
Surprise	0.95	0.94	0.89

F1-score refers harmonic mean of model's precision and recall. A good F1-score refers that the model predicts less false positives and false negatives.

$$F1Score = 2 \times Recall \times Precision\ Recall + Precision \quad (11)$$

Eq. 9, 10, and 11 are applied to predict results obtaining from models to calculate the precision, recall and F1-score, respectively. Tables V, VI and VII contain class wise precision, recall and F1-score of testing data obtained from DenseNet-201, InceptionResNetV2 and VGG-16, respectively.

TABLE VIII. CONFUSION MATRIX OF DENSENET-201 MODEL WITH EYE EMOTION DATASET

Expressions	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	24	1	1	0	1	0
Disgust	3	10	0	2	0	0
Fear	0	0	23	0	0	0
Happy	0	1	0	29	1	3
Sad	0	1	0	1	19	0
Surprise	0	0	0	0	0	18

B. *Error Analysis*

Table VIII contains the confusion matrix of the DenseNet-201 model with the eye emotion dataset. Confusion matrix or error matrix is a kind of mode's prediction summary refers the classification problems. Overall, 89.13% accuracy is achieved to classify the human emotion from the eye and its surrounding features. Here in Table VII, the actual 'Disgust' class several times is predicted as 'Anger' and 'Happy'. There are some common features in the Disgust, Anger, and Happy classes shown in Fig. 9. These features are important parameters for all of the classes [15]. Sometimes, these common features are so much prominent to other features of the classes that the model confused with the actual class to other classes.

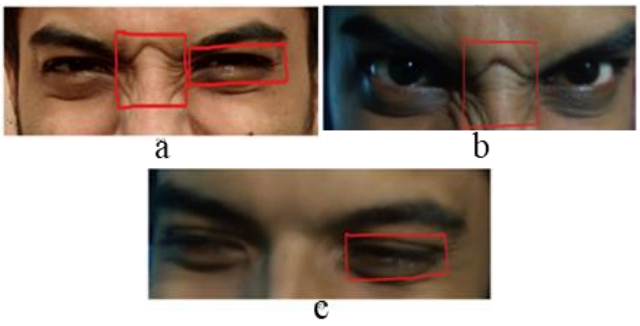

Fig. 9. Feature Similarities among (a) Disgust, (b), Angry, and (c) Happy Classes.

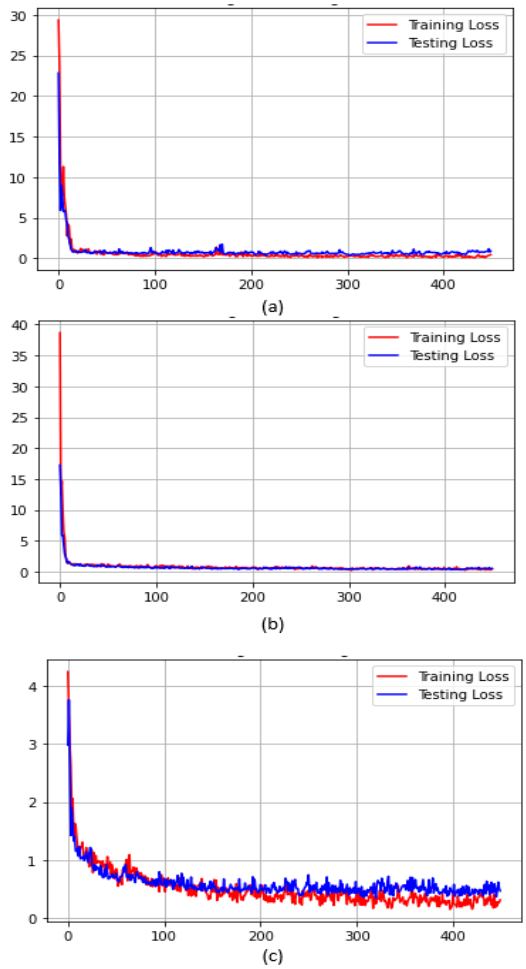

Fig. 10. Loss Graph of DenseNet-201(a), InceptionResNet-V2 (b) and VGG-16(c).

VIII. Comparative Analysis

As mentioned earlier (Section II), many studies have been done to find out the human emotion recognition using machine learning, deep learning, and other techniques based on data (s) from different facial expressions of people.

TABLE IX. Comparison of the Prior Methods with the Proposed Method

Studies	Methods	Accuracy	Datasets
G Verma, et al.[10]	Hybrid CNN	97.07% of FER2013 and 94.12% of JAFFE	FER2013, JAFFE
S. A. Fatima et al. [11]	Mini-Xception, CNN	95.60% from Mini-Xception	FRR2013, Real Time Data
N. Jain et al. [13]	Hybrid Convolutional-RNN	94.91% of JAFFE, 92.07% of MMI	JAFFE, MMI
A. H. Mary et al. [14]	Deep CNN	92.81%	Universal and Personal data
Xuanyu He et al. [17]	CNN	64.6%	Universal data, ArtPhoto, Paintings
S. Palaniswamy et al. [19]	Viola- jones, Active Shape Model, AdaBoost	96%	CMU-MultiPIE, Survey data

Some notable information of the previous studies is shown in Table IX. Many authors have used various methods such as Hybrid CNN, Deep CNN, Hybrid RNN, Viola Jones model and they showed accuracy: 97%, 92%, 94.91%, and 96% respectively. It is worth noting that these studies considered the entire face to recognize the human emotions. However, this study mainly focuses on the data from the eyes and its surrounding areas only. The whole face detection system can identify a human face present in an image/video – it cannot identify that person, but the eyes and its surrounding area can give more precision and accuracy to recognize the person. Subsequently, in the proposed approach we achieved better accuracy 95.3% using Inception ResNet-V2 and the outcomes are comparable with the previous studies in terms accuracy and precision.

IX. Conclusion

Nowadays, automated emotion recognition from facial expressions has become a challenging topic of computer vision but we focus specifically on eye expression of facial expression [38]. We proposed a customized model of deep neural network architecture for eye expression. It takes eye images as input then classifies them into either of six eye expressions: happiness, sadness, anger, disgust, fear, surprise. To get higher accuracy we have trained our dataset with pre-trained weights and used a confusion matrix to analyze the prediction. Our top accuracy rate is 91.78% in DenseNet-201 and contains less overfitting of 0.0265.

X. Limitation of the Study and Future Direction

Challenges like partial occlusions, facial incompleteness, the pose of the face, invariance to pose, poor image quality, continuously changing emotions, backlight, illumination variation, and many additional factors in the real-time detection will be under our investigation and further can be explored our in future studies to improve the recognition rate [19]. We strive to improve and develop our proposed system in several directions. Some other primary and secondary facial expressions will be added to our dataset and other advanced deep learning models with superior learning capabilities, better performance, shorter operation time, and higher classification accuracy can be implemented for testing and comparing the accuracy and ensuring more accurate recognition of emotions.

References

[1] V. S. Johnston, Why we feel: The science of human emotions. Perseus Publishing, 1999.

[2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” Insights Imaging, vol. 9, no. 4, pp. 611–629, 2018.

[3] M. M. Hassan, S. Huda, J. Yearwood, H. F. Jelinek, and A. Almogren, “Multistage fusion approaches based on a generative model and multivariate exponentially weighted moving average for diagnosis of cardiovascular autonomic nerve dysfunction,” Inf. Fusion, vol. 41, pp. 105–118, 2018.

[4] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, “Human emotion recognition using deep belief network architecture,” Inf. Fusion, vol. 51, pp. 10–18, 2019.

[5] G. Lee, M. Kwon, S. K. Sri, and M. Lee, “Emotion recognition based on 3D fuzzy visual and EEG features in movie clips,” Neurocomputing, vol. 144, pp. 560–568, 2014.

[6] E. Kanjo, E. M. G. Younis, and N. Sherkat, “Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach,” Inf. Fusion, vol. 40, pp. 18–31, 2018.

[7] Z. Liu et al., “A facial expression emotion recognition based human-robot interaction system,” 2017.

[8] A. S. Al-Waisy, R. Qahwaji, S. Ipson, S. Al-Fahdawi, and T. A. M. Nagem, “A multi-biometric iris recognition system based on a deep learning approach,” Pattern Anal. Appl., vol. 21, no. 3, pp. 783–802, 2018.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[10] G. Verma and H. Verma, “Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions,” Rev. Socionetwork Strateg., vol. 14, no. 2, pp. 171–180, 2020.

[11] S. A. Fatima, A. Kumar, and S. S. Raoof, “Real Time Emotion Detection of Humans Using Mini-Xception Algorithm,” in IOP Conference Series: Materials Science and Engineering, 2021, vol. 1042, no. 1, p. 12027.

[12] T. Wiatowski and H. Bölcskei, “A mathematical theory of deep convolutional neural networks for feature extraction,” IEEE Trans. Inf. Theory, vol. 64, no. 3, pp. 1845–1866, 2017.

[13] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, “Hybrid deep neural networks for face emotion recognition,” Pattern Recognit. Lett., vol. 115, pp. 101–106, 2018.

[14] A. H. Mary, Z. B. Kadhim, and Z. S. Sharqi, “Face Recognition and Emotion Recognition from Facial Expression Using Deep Learning Neural Network,” in IOP Conference Series: Materials Science and Engineering, 2020, vol. 928, no. 3, p. 32061.

[15] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, “Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks,” arXiv Prepr. arXiv1410.8586, 2014.

[16] Md. Ashiqul Islam; Md. Nymur Rahman Shuvo; Muhammad Shamsojjaman; Shazid Hasan; Md. Shahadat Hossain; Tania Khatun, “An Automated Convolutional Neural Network Based Approach for

Paddy Leaf Disease Detection," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 1, 2021, doi: 10.14569/IJACSA.2021.0120134.

[17] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," Neurocomputing, vol. 291, pp. 187–194, 2018.

[18] A. Mahmood, S. Hussain, K. Iqbal, and W. S. Elkilani, "Recognition of facial expressions under varying conditions using dual-feature fusion," Math. Probl. Eng., vol. 2019, 2019.

[19] S. Palaniswamy and S. Tripathi, "Emotion Recognition from Facial Expressions using Images with Pose, Illumination and Age Variation for Human-Computer/Robot Interaction.," J. ICT Res. Appl., vol. 12, no. 1, 2018.

[20] C. Reddy, U. Reddy, and K. Kishore, "Facial Emotion Recognition Using NLPCA and SVM.," Trait. du Signal, vol. 36, no. 1, pp. 13–22, 2019.

[21] D. G. R. Kola and S. K. Samayamantula, "A novel approach for facial expression recognition using local binary pattern with adaptive window," Multimed. Tools Appl., pp. 1–20, 2020.

[22] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," Procedia Comput. Sci., vol. 108, pp. 1175–1184, 2017.

[23] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," IEEE Trans. Cybern., vol. 49, no. 3, pp. 1110–1122, 2018.

[24] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," Appl. Sci., vol. 7, no. 1, p. 110, 2017.

[25] A. Kartali, M. Roglić, M. Barjaktarović, M. Đurić-Jovičić, and M. M. Janković, "Real-time Algorithms for Facial Emotion Recognition: A Comparison of Different Approaches," in 2018 14th Symposium on Neural Networks and Applications (NEUREL), 2018, pp. 1–4.

[26] V. Pandit, M. Schmitt, N. Cummins, and B. Schuller, "I see it in your eyes: Training the shallowest-possible CNN to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time," Inf. Process. Manag., vol. 57, no. 6, p. 102347, 2020.

[27] J.-J. Guo, R. Zhou, L.-M. Zhao, and B.-L. Lu, "Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 3071–3074.

[28] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 933–936.

[29] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 223–232.

[30] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 83–92.

[31] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 47–56.

[32] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 229–238.

[33] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in 2015 3rd IAPR Asian conference on pattern recognition (ACPR), 2015, pp. 730–734.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition (2015)," arXiv Prepr. arXiv1512.03385, 2016.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2017, vol. 31, no. 1.

[37] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," Pattern Recognit., vol. 61, pp. 610–628, 2017.

[38] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: taxonomy, review and current challenges," Sensors, vol. 20, no. 8, p. 2384, 2020.