Journal of Theoretical and Applied Information Technology
30th April 2024. Vol.102. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP HUMAN FACIAL EMOTION RECOGNITION: A TRANSFER LEARNING APPROACH USING EFFICIENTNETB0 MODEL

R. ANGELINE[1], A. ALICE NITHYA[2]

[1]Research Scholar, Department of Computational Intelligence, SRMIST, Kattankulathur, Chennai, Tamilnadu, India
[2]Associate Professor, Department of Computational Intelligence, SRMIST, Kattankulathur, Chennai, Tamilnadu, India
E-mail: [1]ar1501@srmist.edu.in, [2]alicenia@srmist.edu.in

ABSTRACT

Because of its potential usefulness, Facial Emotion Recognition (FER) has become one of the computer vision field's fastest-growing applications. One of the main ways to communicate appropriately with people is through facial expressions. Communication success largely depends on one's capacity to read others' facial expressions. Finding the emotional states connected to varied facial expressions is the main objective of FER. In this research, different emotions are analysed and classified into eight categories: anger, contempt, disgust, fear, happiness, neutrality, sorrow, and surprise, using a CNN-based transfer learning approach. By employing 33,000+ accurately re-annotated images from popular datasets like FER2013, KDEF, and CK+ to train our model using a transfer learning strategy. EfficientNetB0, pre-trained on the ImageNet dataset, is used as the base model in this paper. The above model is fine-tuned and validated after performing Test Time Augmentation (TTA), achieving a training accuracy of 87%.

Keywords: *Facial Emotion Recognition, Anger, Contempt, Disgust, Fear, Happiness, Neutrality, Sorrow, Srprise, CK+, Test Time Augmentation.*

1. INTRODUCTION

Face expressions are essential in comprehending human emotions. These have a significant impact on a person's social interactions as well. Understanding these expressions on the face is crucial for non-verbal communication, which makes up around two-thirds of all human interaction.

Novel hand-over-face gesture-based emotion recognition technique [1] used extra hand movements and deep learning. The Convolutional Neural Network (CNN) used in this study eliminates the need for manual feature extraction and automatically extracts additional class-specific features. A Recurrent Neural Network (RNN) is utilized to learn and classify these ever-more complex emotional characteristics. In addition to the basic emotions, the research area has included more complex emotions in our advised course of action, such as self-assurance, decision-making ability, dread, and guilt.

Recurrent neural networks (RNNs) are linked with convolutional neural networks (CNNs), which are used to assess and recognize continuous facial expressions over time, to improve the traditional image identification technique [2]. The deficiencies are examined in this study compared to the current deep convolution neural network optimization for facial expression recognition [3]. The fer2013 data set is used to train the convolutional network. The approach used in this study is successful at identifying facial emotions.

A convolutional neural network (CNN) [4], which employs four consecutive sets of layers as well as a loss function, is used in this study to classify each facial image into one of the seven categories of human emotions. The data set for the 2013 Facial Emotion Recognition Challenge (FER2013) on Kaggle consists of 35,887 grayscale

portraits of people with 48 by 48-pixel faces, each labeled with a single emotion category. In about 64% of cases, the model is accurate.

To create geometric-based features for films from the BioVid Emo database, RNN [5] detects the face in videos and extracts local characteristics (landmarks) to distinguish between a set of five emotion expressions: amusement, anger, disgust, fear, and sadness.

The majority of speech emotion research uses prosodic characteristics. It was discovered in 2018 when examining the emotion-specific features of the FER system [6] using a PCA technique. Energy, duration, pitch, and their derivatives are considered to be strong correlates of emotion. It was suggested [5] that performance evaluation in educational contexts shows that the factor graph-based FER model can exceed the current baselines.
The traditional method of Facial Emotion Recognition (FER) may be broken down into three distinct phases. [7] At first, a picture of a face is recognized, and facial components of the face are identified from the area. It is possible to have an eye, brows, or a nose and a mouth on the front. Secondly, various regions of the face will be analyzed in order to extract features. The last step is to train a classifier on the training data before using it to produce labels for the various emotions.

Automatic recognition of elearners cognitive state, interaction and the feedback was developed as a model [8] with the framework to identify the basic emotions of happy, sad, surprise, fear and angry in facial expression analysis. In this facial expression analysis, the facial emotions are categorised into positive and negative. PERcentage CLOSure of eye (PERCLOS) and saccadic parameters are used to estimate the alertness level. The ocular measurement device provides various aspects of eye alertness which is useful to measure the alertness level.

New approaches to transfer learning (TL) make use of information analyzed from one area of study and apply it to another. When it comes to expressing emotion, it is common for individuals to exhibit common traits. Angry people, for example, may speak more loudly and have more extreme facial expressions. Fear is often indicated in a quieter voice and a higher heart rate [9].

Facial emotion recognition using Transfer Learning in the Deep CNN [10] on 10-fold cross-validation with DenseNet-161 on datasets of KDEF and JAFEE with 96.51% and 99.52% accuracy respectively. These offer insightful data that can be examined to comprehend how a person's surroundings affect his feelings. The most noticeable aspects of a person's face are their eyes, lips, jaw, and eyebrows, which make it easier to categorize their feelings into eight fundamental emotion groups: anger, disgust, sadness, happiness, fear, surprise, contempt, and neutral [11,12].

A recent assessment of FER and classification found that the first framework uses a modified approach dubbed "Viola-Jones Face Detection"[13] to locate faces. Emotion classification is then accomplished by utilizing a variety of classifiers. Afterward, the features can be extracted using [14]"Zernike moments," [15]"DCT transform," or [16] "LBP."

The similarity across different emotional datasets may be due to the presence of certain emotion-specific features being satisfied by many individuals. As a result, transfer learning techniques may learn common patterns related to emotions and can be used across domains to recognize emotions in datasets with few or non-labeled samples [20].
Various other methods use convolutional neural networks' facial emotion recognition. The technique employs a two-section convolutional neural network (CNN) to extract the face component vector after first removing the background from the image.

2. EXPERIMENTAL METHODOLOGY:

2.1 Sections and Subsections

In the fields of computer vision and artificial intelligence, recognizing human emotions is a difficult task. Deep learning [21-23] has made it possible to classify emotions from face photos with exceptional accuracy under realistic situations, nevertheless, in the last ten years.

CNN's trained with millions of images have been used to solve this issue[17], but the process is time-consuming and requires a lot of data. A pre-trained model may be repurposed, avoiding the need for a large amount of data, and maximizing processing resources. The most significant issue is when the intended data collection needs to be bigger to be challenging to work with. Overfitting may be an issue in many circumstances, and data augmentation may only sometimes be enough to fix it.

EfficientNetB0, an emotion recognition system pre-trained on ImageNet [18], is used in this study to propose a transfer learning method that focuses on salient face areas. Standard datasets like FER, CK-Plus, and KDEF were used, which provided about 33,000 images. The accuracy was improved with corrective re-annotation of images

2.2 Transfer learning

In general, the term "transfer learning" describes a process wherein the expertise obtained in one problem domain is applied to problems in another domain. Operationally, it means that the model developed for one task is used for a different task.

Transfer learning in deep learning offers the advantage of cutting down on neural network model training time and lowering generalization error [9-10].

The pre-trained model's weights are utilized as the starting point for the process of training the new problem. When the first linked problem contains a lot more marked data than the actual problem, this could be helpful.

The primary model in Figure 1 using EfficientNetB0, was pre-trained using an image database called ImageNet, containing more than 14 million annotated images according to the WordNet hierarchy.

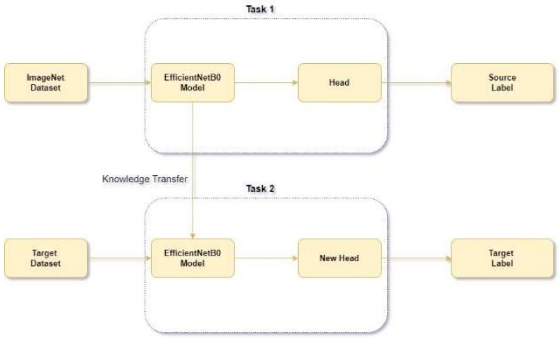

Figure 1 Transfer Learning Model using EfficientNetB0

2.3 Model architecture

One of EfficientNet's eight models, EfficientNetB0, serves as the foundation model in the transfer learning strategy for facial emotion recognition (FER). This model suggests a brand-new scaling technique that uses a straightforward yet incredibly potent compound coefficient to equally scale all depth, breadth, and resolution variables. This model attained modern 84.3 percent accuracy on the ImageNet dataset. When compared to the best existing convolutional networks, it is also quicker and smaller. Figure 2 shows the accuracy graph using EfficientNetB0.

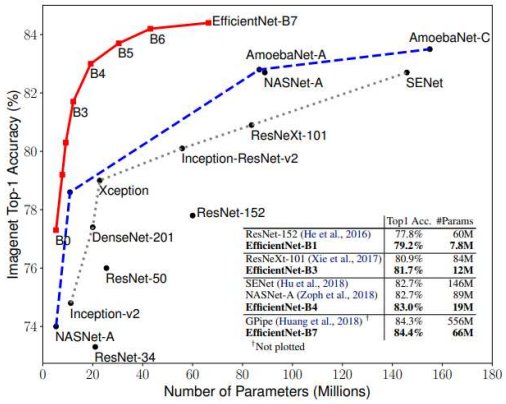

Figure 2 Accuracy graph using EfficientNetB0

The architecture of the EfficientNetB0 consists of a total of 237 layers. Its main building block is mobile inverted bottleneck MBConv (paper link) with different settings to which squeeze and excitation optimization and swish activation are added.

Table 1 Channel output for stage I, layers Li, input resolution (Hi, Wi)

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	14 × 14	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

In Table 1, each row describes stage I with Li layers, with input resolution (Hi, Wi) and output channels Ci.

2.3.1 EfficientNet- MBConv Block:

Data and arguments are the first things that the MBConv block needs. Data is output from the final layer. An MBConv block may use attributes like input and output filters, expansion and squeeze ratio, and other attributes that are included in a block argument. Since the inner blocks are more expansive and the linked blocks are narrower, the results obtained by making the layer wider by adding more channels.

Following the expansion, depthwise convolution is carried out using the block parameter's by giving kernel size. In this step, the global average pooling is used to extract global features and se ratio to squeeze channel numbers. The output filters indicated in the argument block are produced using convolution after the se block is complete.

2.3.2 Swish Activation:

The linear and sigmoid functions combine to form the swish activation function. The research demonstrates that on deep models across a variety of difficult datasets, swish typically performs better than ReLU.

f(x) = x · sigmoid(x).

2.3.3 Squeeze and Excitation Block:

When constructing the feature maps for CNN, the network equally weights all of its channels.The squeeze and excite Block aim to alter this strategy by gradually adding weight to each channel. By condensing the feature maps to a single numerical value, they first gain a broad comprehension of each channel. A vector of size n is produced as a result, where n is the number of convolutional channels. It is then input into a two-layer neural network, which produces an output vector of the same size. These n values can now be applied to the original feature maps as weights, sizing each channel according to its significance. Figure 3 represents the pictorial representation of the ResNet Module and SE-ResNet Module.

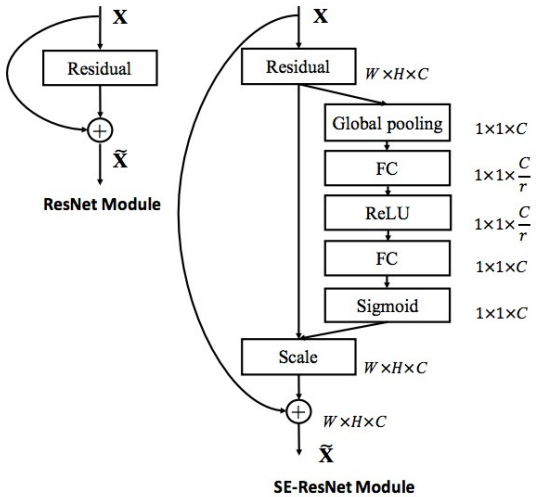

Figure 3 ResNet and SE-ResNet module representation

2.3.4 Fine-Tuning Efficientnetb0:

The EfficientNetB0 has a total of 237 layers in which all the pre-trained are kept layers frozen and added three Fully Connected layers with 50% dropout and ReLU activation and an output softmax layer at the end. After using adam optimizer, categorical cross-entropy loss function, and various other methods, by achieving an accuracy of 87%.

2.4 Loss function:

The loss function used in this model is Categorical Crossentropy, which calculates the loss by computing the following formula:

$$Loss = -\sum_{i=1}^{outputsize} y_i \,.\, log\hat{y}_i \quad \text{------(1)}$$

2.5 Methods:

2.5.1 Data Augmentation:

Data are examined and tested several methods that were frequently employed in FER publications and found that horizontal mirroring, 10-degree rotations, ten percent image zoom, and ten percent horizontal/vertical shifting gave us the best results.

2.5.2 Test-Time Augmentation:

For each test image, multiple versions that are different from the original are created. The original images are cropped, flipped, rotated, and zoomed to form new images. TTA aims at boosting the model accuracy in the inference stage.

3 RESULT ANALYSIS

3.1 Dataset

In this paper, the dataset consists of correctly re-annotated images from industry standard datasets like FER, CK+, and KDEF. This dataset was created and open-sourced by Sudarshan Vaidya[19], who found that several images from these standard datasets were mis-categorized and manually re-annotated in order to increase the model accuracy.

The dataset contains 33,000 + images with eight different emotion categories such as anger, contempt, fear, disgust, happiness, neutrality,

sadness, and surprise. Every image has 224 x 224 pixels containing grayscale human faces in PNG format. Figure 1 of Re-Annotated images gives the original and revised manually annotated images.

Figure 1 Re-Annotated images

Image	Original FER Annotation	Revised Manual Annotation
	Anger	Happiness
	Calmness	Sadness
	Fear	Neutral
	Happiness	Neutral
	Sadness	Happiness

3.2 Result Analysis:

The result analysis of the Sequential Model with layer type and output shape for corresponding parameters is given in figure 4

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
efficientnetb0 (Functional)  (None, 2, 2, 1280)        4049571
_________________________________________________________________
dropout (Dropout)            (None, 2, 2, 1280)        0
_________________________________________________________________
flatten (Flatten)            (None, 5120)              0
_________________________________________________________________
batch_normalization (BatchNo (None, 5120)              20480
_________________________________________________________________
dense (Dense)                (None, 32)                163872
_________________________________________________________________
batch_normalization_1 (Batch (None, 32)                128
_________________________________________________________________
activation (Activation)      (None, 32)                0
_________________________________________________________________
dropout_1 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_1 (Dense)              (None, 32)                1056
_________________________________________________________________
batch_normalization_2 (Batch (None, 32)                128
_________________________________________________________________
activation_1 (Activation)    (None, 32)                0
_________________________________________________________________
dropout_2 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_2 (Dense)              (None, 32)                1056
_________________________________________________________________
batch_normalization_3 (Batch (None, 32)                128
_________________________________________________________________
activation_2 (Activation)    (None, 32)                0
_________________________________________________________________
dense_3 (Dense)              (None, 8)                 264
=================================================================
Total params: 4,236,683
Trainable params: 589,480
Non-trainable params: 3,647,203
```

Figure 4 Sequential model parameter results

This model summary shows the fine-tuning of EfficientNetB0 with three fully connected layers, the total trainable and non-trainable parameters, and the shape of the output

Table 4: Comparison of accuracy, Loss, Val_accuracy, and Val_loss before and after corrective re-annotation

Result	Accuracy	Loss	Val_accuracy	Val_loss
Model before Corrective re-annotation	0.8571	1.8078	0.8571	1.8099
Model after Corrective re-annotation	0.8750	1.8380	0.8750	1.8350

From the above table, it is clearly proven that by corrective re-annotation, the accuracy of the model has improved from 0.8571 to 0.8750. Thus corrective re-annotation of the dataset has improved the performance of the model.

3.3 Evaluation Metrics:

3.3.1 Accuracy:

The model's accuracy score is the number of accurate predictions divided by the overall number of projections.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Where TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative respectively.

3.3.2 Area under the Curve (AUC):

The ability of a model to distinguish between classes is measured by its AUC. The better a model makes predictions, the higher its AUC score.

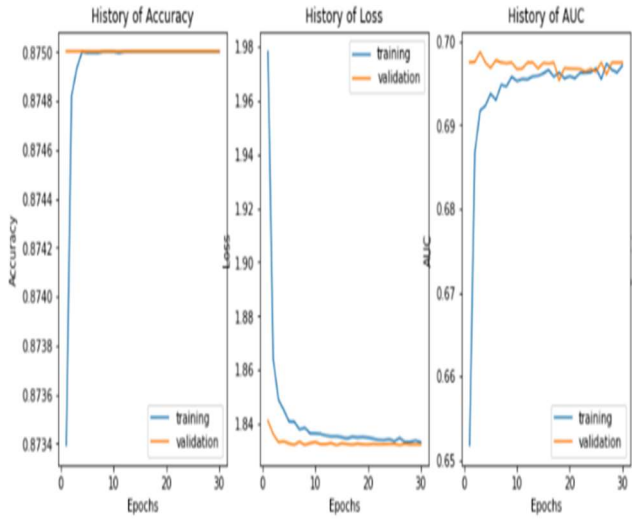

Figure 5 Line Graph for proposed Model Training Accuracy, Loss, and AUC

From the Figure 5 line graph, it is clearly evident that as the number of epochs increases, the accuracy of the model increases and gets saturated after some point. The final training accuracy of the model is 0.8750, mainly due to the re-annotation of the dataset. The training loss has fallen from 1.9 to 1.8380, and the validation loss is 1.8350. This model also has higher training and validation AUC score of nearly 0.7 when compared to 0.64 of the previous models without re-annotation.

4 CONCLUSION AND FUTURE WORK

One of the key challenges in the creation of a robust emotion detection system is the ability to generate well-generalized models with little data. Transfer learning has been studied as a potential solution, but there are still issues with it, such as the need to relinquish previously acquired knowledge, that need to be resolved. The EfficientNetB0 CNN model, which is based on transfer learning, is used in this study to recognise and classify a variety of human emotions. The article's key contribution would be its accuracy in identifying different emotional states using an EfficientNet-B0 CNN model. The model's final training accuracy is 0.8750, mostly as a result of the re-annotation of the dataset. In order to form reliable automatic recognition systems, work must be done using RGB datasets created under uncontrolled conditions, deep neural networks as emotion classifiers, compound emotions, micro-expressions, and multimodal behavioural systems such as body movements, voice, and others. Deep neural networks are also needed to classify emotions.

REFERENCES:

[1] N. Naik and M. A. Mehta's, "Hand-over-Face Gesture based Facial Emotion Recognition using Deep Learning," published in 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018, pp. 1-7, doi: 10.1109/ICCSDET.2018.8821186.

[2] S. Lin, Y. Tseng, C. Wu, Y. Kung, Y. Chen and C. Wu's, "A Continuous Facial Expression Recognition Model based on Deep Learning Method," published in 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2019, pp. 1-2, doi: 10.1109/ISPACS48206.2019.8986360.

[3] L. Liu's, "Human Face Expression Recognition Based on Deep Learning-Deep Convolutional Neural Network,"published in 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2019, pp. 221-224, doi: 10.1109/ICSGEA.2019.00058.

[4] A. Deopa, A. Sinha, A. Prakash and R. K. Sinha's, "Facial Expression Recognition using Convolutional Neural Network and SoftMax function on Captured Images," published in International Conference on Communication and Electronics Systems (ICCES), 2019, pp. 273-279, doi: 10.1109/ICCES45898.2019.9002524.

[5]Amr Mostafa, Mahmoud I.Khalil, Hazem Abbas, 'Emotion Recognition by Facial Features usingRecurrent Neural Networks(RNN), published in the 2019 International Symposium

on Intelligent Signal Processing and Communication Systems (ISPACS) .

[6] Kartika Candra Kirana, Slamet Wibawanto, Heru Wahyu Herwanto's, " Exploring Emotion Specific Features for Emotion Recognition System using PCA Approach", published in International Conference on Solid State Devices - 2018

[7] Mar Saneiro, Olga C. Santos, Sergio Salmeron-Majadas, and Jesus G. Boticario's," Towards Emotion Detection in Educational Scenarios from Facial Expressions and Body Movements through Multimodal Approaches", published in Hindawi Publishing Corporation Scientific World Journal Volume 2014, Article ID 484873.

[8] S. L. Happy, A. Dasgupta, P. Patnaik, and A. Routray's, "Automated alertness and emotion detection for empathic feedback during E-learning," published in IEEE conference in the year 2013. T4E 2013, pp. 47–50, 2013.

[9] Kexin Feng and Theodora Chaspari's, "A Review of Generalizable Transfer Learning in Automatic Emotion Recognition", Front. Comput. Sci., 2020.

[10] M. A. H. Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN", Electronics 2021, 10, 1036. https://doi.org/10.3390/electronics10091036.

[11] Yosinski, J, Clune, J, Bengio, Y. Lipson's research on" How transferable are features in deep neural networks?? In Advances in Neural Information Processing Systems 27".

[12] Rawat, W.; Wang, Z's," Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review of Neural network Computation". 2017, 29, 2352–2449.

[13] A. Rizqullah, N. F. A. Hakim, S. A. T. A. Azhima and I. Kustiawan, "Face Recognition Based on Viola-Jones Algorithm as Dataset for Image Classification," *2021 3rd International Symposium on Material and Electrical Engineering Conference (ISMEE)*, 2021, pp. 295-298, doi: 10.1109/ISMEE54273.2021.9774251.

[14] T. Theodoridis, K. Loumponias, N. Vretos and P. Daras, "Zernike Pooling: Generalizing Average Pooling Using Zernike Moments," in *IEEE Access*, vol. 9, pp. 121128-121136, 2021, doi: 10.1109/ACCESS.2021.3108630.

[15]Z. N. Nacer, A. Zergainoh and A. Merigot, "Global discrete cosine transform for image compression," *Proceedings of the Sixth International Symposium on Signal Processing and its Applications (Cat.No.01EX467)*, 2001, pp. 545-548 vol.2, doi: 10.1109/ISSPA.2001.950201.

[16] S. Karanwal, "Improved LBP and Discriminative LBP: Two novel local descriptors for Face Recognition," *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, 2022, pp. 1-6, doi: 10.1109/ICDSIS55133.2022.9915933.

[17] S. Giri *et al*., "Emotion Detection with Facial Feature Recognition Using CNN & OpenCV," *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2022, pp. 230-232, doi: 10.1109/ICACITE53722.2022.9823786.

[18] H. Yin, C. Yang and J. Lu, "Research on Remote Sensing Image Classification Algorithm Based on EfficientNet," 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022, pp. 1757-1761, doi: 10.1109/ICSP54964.2022.9778437.

[19] Sudharshan Vaidhya; Detecting Human Emotions - Facial Expression Recognition; medium.com.

[20] A. Poulose, J. H. Kim and D. S. Han, "Feature Vector Extraction Technique for Facial Emotion Recognition Using Facial Landmarks," 2021 International Conference on Information and Communication Technology Convergence (ICTC), 2021, pp. 1072-1076, doi: 10.1109/ICTC52510.2021.9620798.

[21] Rakesh, Geethalakshmi, and Vani Rajamanickam. "A Novel Deep Learning Algorithm for Optical Disc Segmentation for Glaucoma Diagnosis." Traitement du Signal 39, no. 1 (2022).

[22] Sivakumari, T., and R. Vani. "Implementation of AlexNet for Classification of Knee Osteoarthritis." In 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp. 1405-1409. IEEE, 2022.

[23] Vani, R., J. C. Kavitha, and D. Subitha. "Novel approach for melanoma detection through iterative deep vector network." Journal of Ambient Intelligence and Humanized Computing (2021): 1-10.