2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC)

Facial Emotional Recognition using Faster Regional Convolutional Neural Network with VGG16 Feature Extraction Model

1st Nagendar Yamsani
School of Computer Science and Artificial Intelligence
SR University
Warangal, Telangana, India
nagendar.yamsani@gmail.com

2nd Muhammed Basim Jabar
National University of Science and Technology
Dhi Qar, Iraq
muhammed.basim.jabar@gmail.com

3rd Myasar Mundher adnan
The Islamic university
Najaf, Iraq
abathermahmood560@gmail.com

4th A. H. A. Hussein
College of Pharmacy, Ahl Al Bayt University
Karbala, Iraq
AH.hussien@abu.edu.iq

5th Subhra Chakraborty
Department of Electronics and Communication Engineering
Nitte Meenakshi Institute of Technology
Bengaluru, India
subhra.chakraborty@nmit.ac.in

2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC) | 979-8-3503-1702-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMNWC60182.2023.10435819

Abstract—**Facial Emotion Recognition (FER) poses a significant challenge when training deep learning (DL) models like Artificial Neural Network (ANN) and machine learning (ML). The primary issue in FER is the detection of the face, the creation of bounding boxes, and the classification of emotional expressions. This paper proposes the use of Faster Regional Convolutional Neural Network (RCNN) for facial recognition and employs a median filter to filter the image dataset. The model leverages the Facial Emotional Recognition (FER 2013) dataset for training and predicting accurate classifications. The Median Filter is used to eliminate noise in the image, such as Salt and Pepper noise, and data augmentation is utilized to generate synthetic data for addressing minor imbalances in classes. Subsequently, the VGG 16 architecture is implemented for precise feature extraction to reduce complexity in the classification process. This paper, achieved accuracy, precision, recall, specificity, and f1-score values of approximately 78.22%, 75.40%, 80.20%, 85.90%, and 71.40%, respectively, when compared to other classification methods like Convolutional Neural Network, ResNet 50, Divide-And-Conquer-Based learning, and Resnet 18.**

Keywords—Artificial Neural Network, Convolutional Neural Network, Deep Learning, Facial Emotion Recognition, Median Filter.

I. INTRODUCTION

Face detection is a fundamental and crucial aspect of computer vision systems and it serves as an essential initial stage in face verification [1]. In feature-based approaches, the initial step involves processing the input image to identify and extract characteristic facial landmark points associated with regions like the eyes, nose, and mouth, among others [2]. Within the realm of artificial intelligence, cognitive computation focuses on developing computational methods to explore various human mental processes, including beliefs, thoughts, opinions, emotions, personality, and sentiments [3]. An efficient algorithm is used to extract and implement facial features, and various improvements are made to enhance the existing algorithm model [4]. Recent research efforts have delved into classifying facial recognition, with examples including the use of Mixture Deep Neural Networks based on double-channel facial images, Speckle-Based Optical Cryptosystems [5,6], Deep Learning Techniques for Face Recognition, and the development of FS-CNN [7,8]. Previous model encounter challenges such as salt and pepper noise images, over fitting, oversampling, imbalance classes, Median filter, data augmentation, Faster RCNN, are used to overcome those problems and contributions are mentioned below.

- The preprocessing involves the use of a median filter to eliminate salt and pepper noise in Fascial Emotional recognition FER 2013 dataset.
- The VGG 16 is used to extract the various characteristic form the image dataset.
- Faster RCNN is utilized to reduce the time complexity and mitigate overfitting in the classification process.

The remaining part of the research can be outlined as follows: Section 2 provides an overview of related studies, Section 3 elucidates the proposed methodology, Section 4 presents the results and comparative analysis of the proposed approach, and finally, Section 5 offers a conclusion for the paper.

II. LITERATURE REVIEW

Dharma Karan Reddy Gaddam et al. [9] developed a human facial emotion detection system using the Resnet50. They employed Resnet50 with a single layer to represent any function. Initially, grayscale images were used for image preprocessing to identify faces in the images. Subsequently, a CNN was utilized to extract features from the preprocessed images, and the experiment was conducted on the FER 2013 dataset. However, the network architecture needed to be made deeper. But, increasing the depth of the network did not yield the desired results and led to the network being prone to oversampling of the data and encountering a vanishing gradient problem.

979-8-3503-1702-2/23/$31.00 ©2023 IEEE

In the past, Dong-Hwan Lee and Jang-Hee Yoo [10] introduced a divide-and-conquer-based (DACL) learning approach for facial expression recognition. They utilized ResNet 18 and MobileNet to detect the facial area in an image and recognize facial expressions. During the preprocessing stage, after the facial area was detected, all images were normalized to the same resolution. The DACL algorithm divided a given problem into two or more subproblems of the same or related type until they were small enough to be solved directly. However, during the model training in the past, it did not halt at a specific training stage, leading to the model encountering the issue of overfitting.

In their study, Sivakumar Depuru et al. [11] developed a Human Emotion Recognition system based on a Deep Convolutional Neural Network (DCNN). They utilized the Emotional Image FER dataset for training the DCNN and employed Artificial Intelligence to recognize facial emotions through various CNN layers. In the preprocessing stage, they performed grayscale conversion and standardized the size of each image. However, the model encountered oversampling issues due to the omission of data augmentation, and the dataset exhibited class imbalance.

In their work, Akash Saravanan et al. [12] crafted a CNN for the task of facial emotion recognition. In this model, the output array's dimensionality was reduced, and a decision was taken to favor a deeper network over a wider one. They incorporated a dropout rate of 0.25 to address overfitting concerns. The CNN model served the dual purpose of feature extraction and human emotion recognition in preprocessed images. Nonetheless, the downsizing applied during the preprocessing stage resulted in a loss of fine-grained detail, leading to the generation of blurry images and presenting an additional challenge in the classification process.

Gaurav Meena et al. [13] introduced a CNN for identifying emotions from facial expressions using a deep learning approach. The CNN model was employed for emotion recognition on the Facial Expression Recognition (FER 2013) dataset. This model framework operated by associating groups of images with common feature combinations. The ReLU was employed to function prompts the neurons to return positive values However, the model did not undergo image enhancement because the dataset contained low-quality and noisy images namely salt and pepper noise.

III. Problem Statement

- The image dataset contained salt and pepper noise, low-quality images, and class imbalances
- The primary issue lies in the CNN model's excessive depth, which leads to overfitting problems.

IV. Objective

- The dataset includes grayscale images with salt and pepper noise. To mitigate this noise, we employed a median filter.
- Data augmentation is employed to address oversampling and mitigate minor class imbalances.
- Fast R-CNN is utilized to classify facial expressions and detect multiple faces within the input images.

V. Proposed Method

This section provides a comprehensive explanation of the proposed method, detailing the stages of Data Collection, Preprocessing, Feature extraction, and Classification. Fig. 1 is representing the flow of the proposed model.

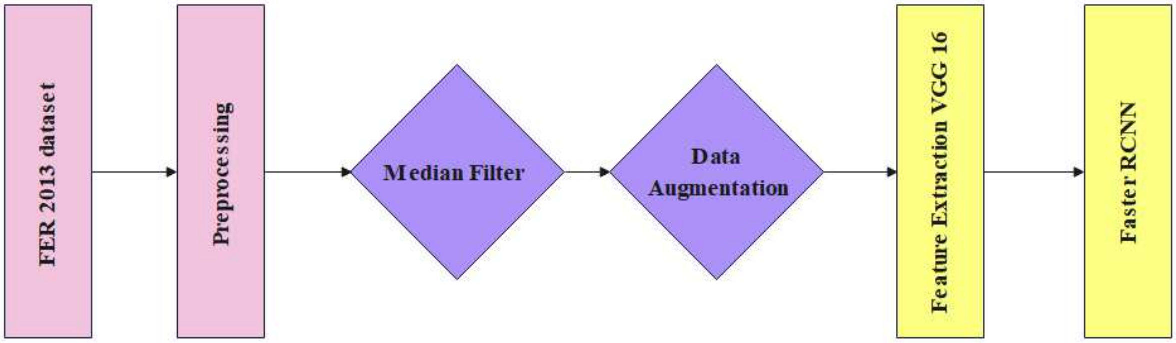

Fig. 1. Flow diagram of Proposed model.

A. Data Collection

The Facial Emotion Recognition (FER2013) dataset [14] comprises grayscale images of faces, each with a resolution of 48x48 pixels. These images have been automatically aligned to ensure that the face is centered and occupies a relatively consistent amount of space in each picture. The objective is to classify each face based on the displayed emotion in the facial expression into one of seven categories: 0 for Angry, 1 for Disgust, 2 for Fear, 3 for Happy, 4 for Sad, 5 for Surprise, and 6 for Neutral. The training dataset contains 28,709 images, while the public test dataset comprises 3,589 images.

B. Preprocessing

The FER 2013 dataset serves as the input for the preprocessing stage. Initially, apply a median filter to all input images. The median filter is effective at eliminating unwanted noise, such as Salt and Pepper Noise, in the image. It is known for its ability to remove low-to-intermediate density impulse noise. The Mean Square Error (MSE) quantifies the average error in relation to the total number of pixels in an image. It is important to note that the definition of MSE does not imply that the noise-removed image introduces additional errors. Instead, it highlights a well-known dissimilarity between the original image and the noise-removed image, indicating critical noise reduction. The formula for calculating MSE (1) is provided by (1).

$$\frac{1}{mn}\sum_{i=0}^{m}\ \sum_{j=0}^{n}[I(i,j)-k(i,j)]^{\,2} \qquad (1)$$

In this context, the original image is denoted as $I\ (i,j)$, and K (i,j) represents the corrupted image. The application of the MF for noise reduction in images yields promising outcomes, particularly in the removal of Salt and Pepper Noise (SPN) in grayscale images. Notably, the mean filter demonstrates superior results compared to those presented in the referenced as [15] paper. Data augmentation is employed to address class imbalance by generating augmented data through methods like rotation, flipping, and color-space transformations applied to the images.

C. Feature Extraction

The feature extraction process in our approach involved the utilization of the VGG16 model. This Convolutional Neural Network (CNN) model integrates various components such as convolutional layers, fully connected layers, pooling, and non-linear operations within the framework of an artificial neural network. The effectiveness of a CNN relies on the specific application and its corresponding training parameters. For our model, we chose the VGG16 architecture [16] for feature extraction. The VGG16 model comprises ten identical deep neural networks equipped with convolutional blocks. We retained the bias and weight information from the original VGG16 model while excluding the fully connected layer. The outputs from these deep neural structures play a pivotal role in constructing the classification network. As depicted in Fig. 2, our feature extraction process leverages ten identical and original VGG16 models, each incorporating convolutional blocks. These models retain the weight and bias from the VGG16 model, excluding the fully connected layer. The outputs from these deep neural structures are then fed into a Multi-Layer Perceptron (MLP) input to construct the classification network.

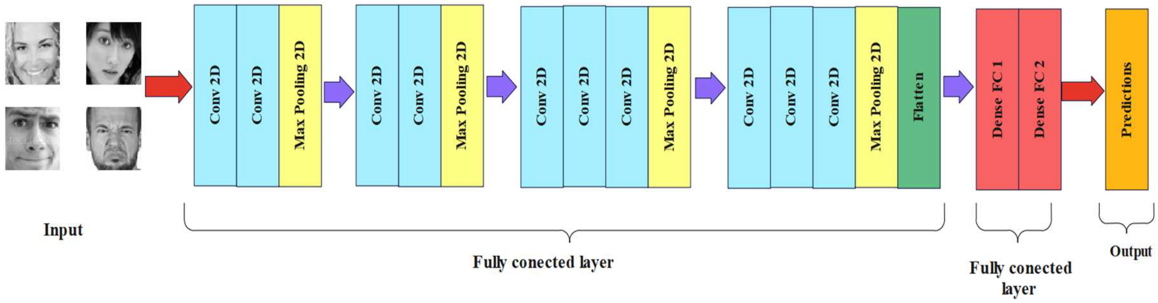

Fig. 2. Architecture of VGG16

D. Faster Region (RCNN) Convolutional Neural Network

The FRCNN stands out as an advanced object detection network, utilizing a two-stage identification approach. In the first stage, it employs a Region Proposal Network to generate a variety of bounding boxes known as region proposals. Within the Faster R-CNN framework, VGG16 is harnessed to extract distinct facial features like eyes, nose, eyebrows, jaw, and mouth. VGG-16 is a pretrained model based on the ImageNet database, serving as the core of the face recognition system's workflow. Images captured are fed into Faster R-CNN for classification using the extracted attribute. This allows the system to detect individuals within the image dataset and recognize their emotions based on the classification results. A color image is used as input in a convolutional network. By enhancing the architecture's complexity in the object identification process, the best performance metrics for Faster RCNN can be achieved. Fig. 3 represent the architecture of Faster R-CNN.

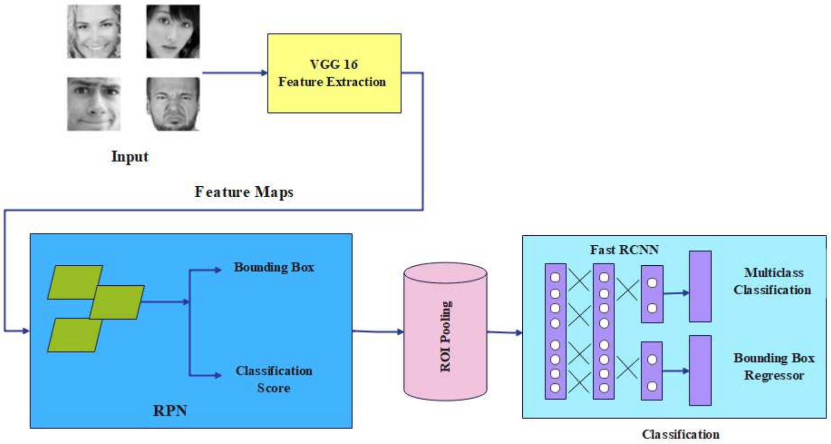

Fig. 3. Architecture of Faster R-CNN

To enhance the effectiveness of personal box shapes, we primarily make adjustments to the pooling layer within the system. The classification layer plays a crucial role in establishing the accurate offset values for the bounding boxes. Consequently, the design of Faster R-CNN significantly contributes to improving the accuracy metrics of the detection method. Notably, Faster R-CNN has been introduced with the aim of reducing the time required for the training process. It achieves this by changing the strategy for region extraction and CNN execution. Instead of the previous approach, which involved finding regions, we now employ a more efficient method. We use a small sliding window that traverses the entire output layer of the convolutional feature map to generate region proposals. This modification has proven to be highly effective. (2) represent the region boundary defined by location coordinates and (3) represent anchor's height.

$$T = \left[\frac{P-P^a}{W^a}, \frac{q^*-q_a}{H_a}, \frac{\log W}{W_a}, \frac{\log H}{H_a}\right] \quad (2)$$

$$T^* = \left[\frac{P-P^*}{W^a}, \frac{q^*-q_a}{H_a}, \frac{\log W^*}{W_a}, \frac{\log H^*}{H_a}\right] \quad (3)$$

The regressors produce a region boundary defined by location coordinates (p, q, H, W). Here eq. (2), pa, qa, Wa, Ha represents the anchor's height, center, and width, while eq. (3) $W *, H *, p *, q *$ correspond to the ground truth bounding box center, width, and height. To calculate the partial derivatives of the loss function with respect to each input variable, we leverage the inverse function of the ROI pooling layer. This inverse function operates by utilizing the argmax switches, allowing us to backtrack through the network and determine the gradients of the loss with respect to the input variables. In (4), for each mini-batch ROI $'p'$ and the output unit b_{pq} accumulate the partial derivative $\frac{\partial l}{\partial b_{pq}}$ when $'i'$ is chosen as the argmax through max pooling. The convolutional operation can be represented as (5) follows:

$$\frac{\partial l}{\partial a_i} = \sum_p \sum_q \; [i = i^*(p.q)] \frac{\partial l}{\partial b_{pq}} \quad (4)$$

$$q\, m^{k+1}, n^{k+1}, f = \sum_{m=0}^{128} \sum_{n=0}^{128} \sum_{f=0}^{N} F_{m,n,f} \times P^{k}_{m^{k+1}} + m, n^{k+1} + m, n \quad (5)$$

Here, within the context, $'p'$ signifies the input image tensor, $'F'$ represents the strainer bank, $'k'$ is the layer number, $'f'$ denotes the filter number, and $'n'$ and $'m'$ refer to the spatial coordinates. $'q'$ stands for the convolution output, and $'N'$ indicates the number of filters. The classification results achieved by the RCNN exhibit the highest accuracy when compared to previous models.

VI. Result

The experimental computer was configured with the following hardware and software components: an NVIDIA GeForce RTX GPU, an Intel Core i7 CPU, 16 GB of RAM, and SSDs for faster data loading and model training. TensorFlow was employed for building, training, and deploying deep neural networks, with Python 3.X used to implement the model on the computer. The code was written using PyCharm software, and Python libraries such as NumPy and OpenCV were utilized to facilitate Python-based operations. The mathematical formula of accuracy, precision, recall, specificity and f1-score are shown in (6), (7), (8), (9) and (10),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

$$F1 - score = 2 \times \frac{Precision \times recall}{Precision + recall} \quad (10)$$

Where, TP, TN, FP and FN represents the True Positive, True Negative, False Positive and False Negatives.

A. Qualitative and Quantitative analysis

This section provides a thorough examination of the VGG16 Method when integrated with the Fast R-CNN Model. The analysis encompasses both qualitative and quantitative aspects, evaluating key performance metrics such as accuracy, precision, recall, specificity, and F1 score. In Table I, you can observe the classification performance of this approach in contrast to several other deep learning methods, including RNN, RCNN, LSTM, and CNN. The comprehensive analysis has been conducted on the FET 2013 Dataset.

TABLE. I. Performance of Classification using FER 2013 Dataset.

Method	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)	Accuracy (%)
RCNN	69.40	67.20	50.30	55.35	60.56
LSTM	75.50	74.65	40.90	68.20	64.81
CNN	74.50	65.30	45.20	70.30	63.82
RNN	60.20	70.25	60.80	75.50	66.68
Proposed Faster R-CNN	75.40	80.20	85.90	71.40	78.22

The results we've obtained clearly demonstrate the superior performance of the proposed model, as indicated by a range of key performance metrics. This model outperforms other classification methods with remarkable scores, achieving an accuracy of 75.40%, precision at 80.20%, recall of 85.90%, specificity reaching 71.40%, and an impressive F1 score of 78.22%.

B. Comparative analysis

This section represents a comparative analysis between the proposed VGG16 model and the Feast RCNN model. The analysis involves performance metrics such as accuracy, precision, recall, specificity, and F1-score, all of which are summarized in Table II. To evaluate the classifier's capabilities, existing results from references [9], [10], and [12]. The proposed method underwent training, testing, and validation using the FER 2013 dataset. The results presented in Table 2 indicate that the VGG16 model in combination with the Feast RCNN outperforms existing methods. Specifically, on the FER 2013 dataset, our model achieved an accuracy of 78.22%, precision of 75.40%, recall of 80.20%, specificity of 85.90%, and F1 score of 71.40%.

TABLE. II. COMPARISON OF PROPOSED MODEL FASTER R-CNN.

Model	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)	Accuracy (%)
CNN-Resnet 50 [9]	65.85	N/A	N/A	60.80	60.70
DACL-Resnet 18 [10]	71.90	N/A	69.50	N/A	77.83
CNN [12]	N/A	70.4	N/A	50.20	68.54
Proposed R-CNN	75.40	80.20	85.90	71.40	78.22

C. Discussion

In this section, let's explore the advantages of the proposed model and shed light on the limitations associated with existing methods. Among the existing techniques, CNN-Resnet50 [9] faced challenges primarily related to increasing the network depth, resulting in suboptimal outcomes and rendering the network vulnerable to overfitting and the vanishing gradient problem. Additionally, the DACL [10] model encountered difficulties during training as it lacked a specific mechanism to halt training at an appropriate stage, consequently leading to oversampling issues stemming from the imbalanced dataset classes. Furthermore, the CNN [12] model suffered from the loss of fine-grained details, which in turn produced blurry images. In contrast, the proposed Faster RCNN model has proven to be an effective solution to these limitations that plagued existing models. It successfully mitigates the risk of overfitting, rectifies class distribution imbalances, enhances the quality of dataset images, and reduces unwanted noise stemming from grayscale image artifacts.

VII. CONCLUSION

Emotions play a fundamental role in human communication and interaction. Developing a facial emotion recognition model is pivotal for enabling computers to comprehend and interpret human emotions. In this paper, propose the utilization of the Faster Region-Convolutional Neural Network (RCNN) for the task of recognizing and classifying facial expressions. Our model is trained on the Facial Expression Recognition (FER2013) dataset. To improve the quality of our input dataset, we initially apply a median filter to reduce both random value impulse noise and fixed value impulse noise. We then leverage the VGG-16 model to extract a wide range of features, effectively addressing concerns such as overfitting, data imbalance, and noise reduction observed in prior models. As a result, our model achieves impressive performance metrics, with an overall accuracy of 78.22%, precision at 75.40%, recall reaching 80.20%, specificity of 85.90%, and an F1 score of 71.40%, surpassing the capabilities of existing models. In future studies, we plan to implement feature selection techniques to further enhance accuracy and reduce the time complexity of the training process.

REFERENCES

[1] Mamieva, D., Abdusalomov, A.B., Mukhiddinov, M. and Whangbo, T.K., 2023. Improved face detection method via learning small faces on hard images based on a deep learning approach. Sensors, 23(1), p.502.

[2] Tome, P., Vera-Rodriguez, R., Fierrez, J. and Ortega-Garcia, J., 2015. Facial soft biometric features for forensic face recognition. Forensic science international, 257, pp.271-284.

[3] Khattak, A., Asghar, M.Z., Ali, M. and Batool, U., 2022. An efficient deep learning technique for facial emotion recognition. Multimedia Tools and Applications, pp.1-35.

[4] Ullah, R., Hayat, H., Siddiqui, A.A., Siddiqui, U.A., Khan, J., Ullah, F., Hassan, S., Hasan, L., Albattah, W., Islam, M. and Karami, G.M., 2022. A real-time framework for human face detection and recognition in cctv images. Mathematical Problems in Engineering, 2022.z

[5] Yang, B., Cao, J., Ni, R. and Zhang, Y., 2017. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE access, 6, pp.4630-4640.

[6] Zhao, Q., Li, H., Yu, Z., Woo, C.M., Zhong, T., Cheng, S., Zheng, Y., Liu, H., Tian, J. and Lai, P., 2022. Speckle-Based Optical Cryptosystem and its Application for Human Face Recognition via Deep Learning. Advanced Science, 9(25), p.2202407.

[7] Fuad, M.T.H., Fime, A.A., Sikder, D., Iftee, M.A.R., Rabbi, J., Al-Rakhami, M.S., Gumaei, A., Sen, O., Fuad, M. and Islam, M.N., 2021. Recent advances in deep learning techniques for face recognition. IEEE Access, 9, pp.99112-99142.

[8] Reshmi, H., Neethu Kuriyakose, B. D. Parameshachari, H. S. DivakaraMurthy, and Arun Jose. "Energy efficiency analysis of compressed sensing video streaming for wireless multimedia sensors." Technology 2, no. 8 (2013).

[9] Gaddam, D.K.R., Ansari, M.D., Vuppala, S., Gunjan, V.K. and Sati, M.M., 2022. Human facial emotion detection using deep learning. In ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications (pp. 1417-1427). Springer Singapore.

[10] Parameshachari, B. D., KM Sunjiv Soyjaudah, and Sumithra KA Devi. "Secure partial image encryption scheme using scan based algorithm." International Journal of advances in Engineering & Technology 6, no. 1 (2013): 264.

[11] Depuru, S., Nandam, A., Sivanantham, S., Amala, K., Akshaya, V. and Saktivel, M., 2022, December. Convolutional Neural Network based Human Emotion Recognition System: A Deep Learning Approach. In 2022 Smart Technologies, Communication and Robotics (STCR) (pp. 1-4). IEEE.

[12] Mehendale, N., 2020. Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2(3), p.446.

[13] Meena, G., Mohbey, K.K., Indian, A., Khan, M.Z. and Kumar, S., 2023. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications, pp.1-22.

[14] Dataset link: https://www.kaggle.com/datasets/msambare/fer2013 (accessed on 17 October 2023).

[15] Jana, B.R., Thotakura, H., Baliyan, A., Sankararao, M., Deshmukh, R.G. and Karanam, S.R., 2023. Pixel density based trimmed median filter for removal of noise from surface image. Applied Nanoscience, 13(2), pp.1017-1028.

[16] Nijaguna, G.S., Babu, J.A., Parameshachari, B.D., de Prado, R.P. and Frnda, J., 2023. Quantum Fruit Fly algorithm and ResNet50-VGG16 for medical diagnosis. Applied Soft Computing, 136, p.110055.