


RESEARCH ARTICLE | SEPTEMBER 05 2023

Deep learning based face recognition and emoji prediction system

Vijayalata Yellasri; Shruti Sharma; Sai Shivani Mudunuri ; Ashlin Deepa Nelson; Misbah Sultana; Bhavani Pale



AIP Conf. Proc. 2754, 070005 (2023)

<https://doi.org/10.1063/5.0167065>



View
Online



Export
Citation

CrossMark

AIP Advances

Why Publish With Us?

-  **25 DAYS**
average time to 1st decision
-  **740+ DOWNLOADS**
average per article
-  **INCLUSIVE**
scope

[Learn More](#)

Deep Learning Based Face Recognition and Emoji Prediction System

Vijayalata Yellasri ^{1, a)}, Shruti Sharma^{1, b)}, Sai Shivani Mudunuri ^{1, c)}, Ashlin Deepa Nelson ^{1, d)},
Misbah Sultana ^{1, e)}, Bhavani Palle ¹

¹Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

a) vijaya@griet.ac.in

b) shrutisharma.22062000@gmail.com,

c) corresponding author: shivanivarma2001@gmail.com

d) rndeepa.pradeep@gmail.com,

e) misbahsultana20ms@gmail.com,

Abstract. The Deep Learning Based Face Recognition and Emoji Prediction System acts as a communication channel in these modern days which is adopted in various fields of science such as psychology, computer science and neuroscience. Emojis have become a way to express emotions and indicate non-verbal cues. The result of this project is a system with the capability of emoji prediction by analyzing the facial expression of the user. The model is an end user application which recognises the expression of the person in the video captured by the web camera. The emoji predicted based on the expression of the person in the video is shown on the screen which changes with the change in the facial expressions of the user. Face detection is achieved by HAAR classification, and the respective emoji is predicted using Mini-Xception architecture which is based on CNN. HAAR Classifier is an important object detection algorithm in machine learning which uses edge and line detection to classify one image from another. The proposed Mini-Xception architecture analyzes visual imagery which has good error detection, efficient processing, and self-developing ability. The model yields an accuracy of 56% while training. This approach has its own advantages of decreasing the parameters and complexity when compared with the basic CNN. The performance of proposed model is compared with CNN. The results of the application may vary because of the differences in the human facial expressions but the model can be trained under different kinds of datasets to acquire accurate predictability.

Keywords: Face Recognition, Emoji Prediction, Image Classification, Deep Learning, HAAR Cascade, CNN

INTRODUCTION

Facial expressions are vital for assessing human behaviour. They are used to infer a person's emotional state. Emotions can depict a lot about the mindset of a person. Studying human behaviour to assess their future actions has become important over years. Technology has enabled scientists to concur facial detection accurately and efficiently. Facial emotion detection from user's expression has become feasible in the last decade due to the advancements in machine learning along with computer vision. Emotion detection has also been researched in various fields such as medical, psychology and neuroscience.

Displaying emojis based on facial expression can make the process of understanding human expression a bit easier. Emojis are being extensively used to communicate over all the internet platforms and associating them with facial expressions can only make the application closer to this generation and their understanding of human behaviour. An emoji related to the user's emotion is displayed. This process involves HAAR cascade [1] for face detection and Mini-Xception CNN model for prediction of the emotion. Facial detection is achieved when the features of the image are extracted with HAAR features. Face region is identified in the image with a window passing all over the region of the image. Face region is identified by creating a bounding box. HAAR cascade is considered to be one of the best ways for frontal face detection.

CNN architecture is often used for object detection. Mini-Xception is a modern CNN architecture with fewer parameters than a traditional CNN architecture. The purpose of the model is to detect emotions like happy, sad, anger, fear, neutral, surprise and disgust from the user's facial expressions. The model is trained with the FER 2013 dataset with labelled images of facial expression. Prediction is made by the proposed model and the corresponding emoji is mapped to the predicted emotion.

RELATED WORK

Theresa Kuntzler et al.,[2] developed a model using FER Tools such as Microsoft Azure Face API, Face++ and Face Reader. All the three tools provide probability of emotions such as neutral, happy, sad, angry etc. The methodology is divided into Face Detection and emoji classification. The human raters' score is computed for each dataset of emotion. For evaluating emotion classification, the algorithm's output is compared to the truth value of emotion. N. Swapna Goud et al.,[3] developed a model using Traditional Convolutional Neural Network (CNN). Several call back functions are called for reducing LOR on plateau and early stopping for reducing any validation loss. The output of this model focuses on dynamic image upload using live webcam.

Ankur Ankit et al.,[4] developed a model which can detect a face using HAAR Cascade and the emoji classification can be done using Support Vector Machine (SVM). In SVM, the features of the images are compared with the trained dataset and after classifying it the emoticon is superimposed on the image. Robert J Henderson et al.,[5] developed a model using Viola Jones classifier for face detection which uses ADA Boosting and cascading classifier for emotion classification. The eigen emotions can be trained in Python and the Tracking.js framework of JavaScript is used for real time images detection from the online webcam application. M.S. Hossain [6] designed an emotion recognition system using Viola Jones algorithm for face detection and KW technique for feature selection which includes null hypothesis and Variance Test. GMM Classifier computes the feature selected and classifies them based on the log-likelihood score. The emoji which gives the maximum score becomes the result of the input.

Abita Devi et al.,[7] classified the emotion recognition system into 3 steps: Face location determination, Feature extraction and emotion classification. Kernel principal component analyses is used for reducing dimensionality of features. In a supervised learning environment, Naïve Baye's classifier is used for emotion classifier and the results of Naive Baye's are compared with the Deep Neural Networks. Daniel Llatas Spiers [8] performed facial emotion detection using Deep learning. Data Augmentation is used for expanding the dataset and the training is done using the parameters: cost function and learning rate to minimize loss and adjust weight and bias accordingly. Ninad Mehendale [9] developed the facial emotion recognition system using CNN by adding the concept of expressional vector (EV). The EV is generated by basic perceptron unit. Expressional vector is nothing but the Euclidian distance between the face parts after normalization. An extra layer of non-convolutional perceptron layer acts as the last stage of this process. Chandra Bhushan Singh et al., [10] developed the model using Support Vector Machine with an extended concept of Infrared active (IR) light to display emotion motion and deformation and when a high grade of resemblance is found, it is said to be detected emotion. Md.Zia Uddin [11] developed Facial Expression Recognition system using Local Directional Rank Histogram, Local Directional Strength Pattern, Generalized Discriminant Analysis and Convolutional Neural Network. LDRHP and LDSP focuses on edge strengths and GDA focuses on depth images. Yingli Tian et al.,[12] developed the model using Viola Jones algorithm for face detection and artificial electric field algorithm for Emotion classification.

The above methodologies have few demerits such as: Naïve Baye's classifier [7],[13],[14] uses independent variables which are impossible to exist in real life, traditional CNN [3],[9] uses a large number of parameters which increases the complexity of the model and Support Vector Machine [4],[10] can't handle large datasets.

PROPOSED APPROACH

The merits of the proposed approach include reduced parameters and reduced complexity which improves efficiency of the model. This is carried out by facial expression recognition. Facial expression recognition [15] contains two processes: Face detection and Emotion recognition. A face should be detected in order to apply emotion detection. For this, a machine learning algorithm exercised to recognize objects in each image, called HAAR Cascade. HAAR

cascade has edge or line detection features and is presented by Viola and Jones in their research paper “Rapid Object Detection using a Boosted Cascade of Simple Features”. This can be accessed via OpenCV. Every frame of the video gets detected using this model. Hence this is considered perfect for frontal face detection. Since HAAR cascade can be used for object detection in real time video and can be applied to develop many applications such as to detect vehicles in a video footage, pedestrian detection from a streaming video etc.

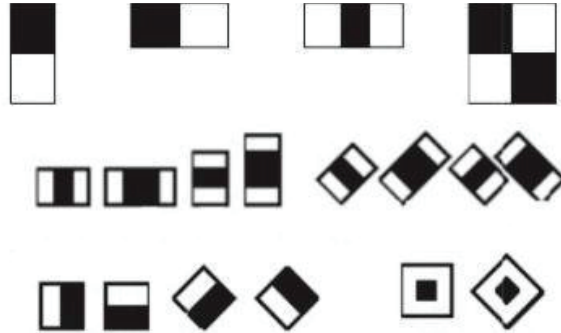


FIGURE 1. HAAR features

Every image contains features that needs to be identified and this is done with the help of HAAR features. HAAR features contain dark areas and light areas as shown in Fig. 1[16]. There are pixels in the features that are dark and light. The pixels that are darker in the feature are valued to be 1 and pixels that are lighter in the feature are valued to be 0. To find out one feature in the image, these pixels in the HAAR features are used. Calculation of the sum of the image pixels of darker and lighter areas is done separately. The difference of these two values is found out to produce a single value. If the calculated value is closer to 1 then the edge is detected. This is done using Integral images. Features that have minimum error rate are selected. To find out if the image has a face or non-face region and classify, these features are used. An image can possess both face and non-face region. The window with non-face region is not needed for further processing and can be ignored. The window is processed in stages, and it passes one stage after another. The window passing through all stages is considered to be a face region. The image with the frontal face is now detected the face region is cropped to reduce the number of frames and maintain uniformity.

The obtained image now is converted to grayscale to reduce computational requirements and simplify the algorithm. Grayscale images only contain one color channel as opposed to three in a color image (RGB). The complexity of grayscale images is lower than that of color images as features relating to brightness, contrast and perspective are being obtained without color. After this, classification [17] needs to be done to predict one emotion out of the seven. This can be done by the Mini-Xception model. Xception algorithm is taken as inspiration. Depth wise separable convolutions along with residual modules [18] are combined in this architecture. Residual modules are used to change the mapping between subsequent layers. These forms learned features which will be the difference between desired features and original features. In a standard convolutional layer, correlations are sought out in both space and depth. This filter will simultaneously consider both spatial and depth dimension. This increases the computations. In the proposed model there is the use of depth wise separable convolutions. They are a combination of depth wise convolution and point wise convolution. Depth wise convolution is carried out independently for each channel spatially and are followed by point wise convolutions which are 1X1 convolution carried across channels. The objective of these layers is to separate spatial from channel correlations [19]. This further reduces the number of computations.

Mini-Xception architecture consists of 4 depth wise separable convolutional layers as shown in Fig. 2. After each convolution batch normalization along with RELU (Rectified Linear Unit) activation function occurs. To obtain a prediction, global average pooling along with soft max activation function are imposed on the final layer [20]. Global average pooling eliminates the need of a fully connected layer and computes the average of features maps generated from every category of the classification task in the final layer and the vector produced is fed into the layer of softmax function. This makes it more robust to deal with spatial input and reduces overfitting as compared to max pooling. A total of 60,000 parameters are present in this architecture which is very less compared to a basic CNN architecture.

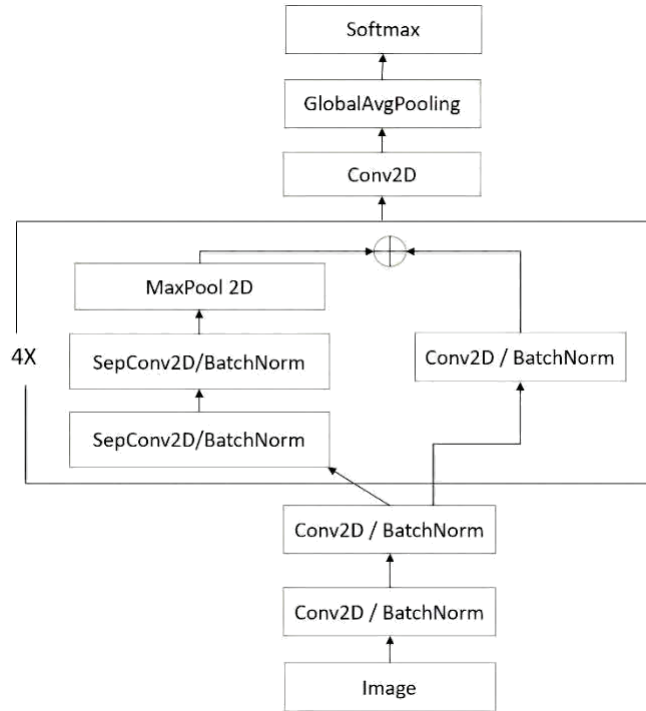


FIGURE 2. Proposed Mini-Xception algorithm approach

Prediction of facial expression is achieved from the final layer of the Mini-Xception model. Using the Mini-Xception algorithm will benefit us by reducing computations and parameters which improves efficiency and accuracy. The following gives the algorithm of proposed approach:

Output: Display Emoji corresponding to the facial expression.

Input: FER 2013 Dataset

1. Acquire the FER 2013 dataset and pre-process the images by scaling them between -1 to 1.
2. The images are split into 80% training set and 20% testing set.
3. Use image augmentation techniques to generate new variations of the image at every epoch. ImageDataGenerator class is utilized to return the images that are transformed with random rotation, horizontal axis flip and horizontal axis shift.
4. The proposed Mini-Xception model is trained with augmented images.
5. The model is tested using test set.
6. During real-time testing, the webcam is used and the face of the user is detected using HAAR cascade model, and the expression is captured. A bounding box appears around the face of the user.
7. Using the trained model of Mini-Xception obtained from step 4, the emoji representing the emotion of the face is predicted.

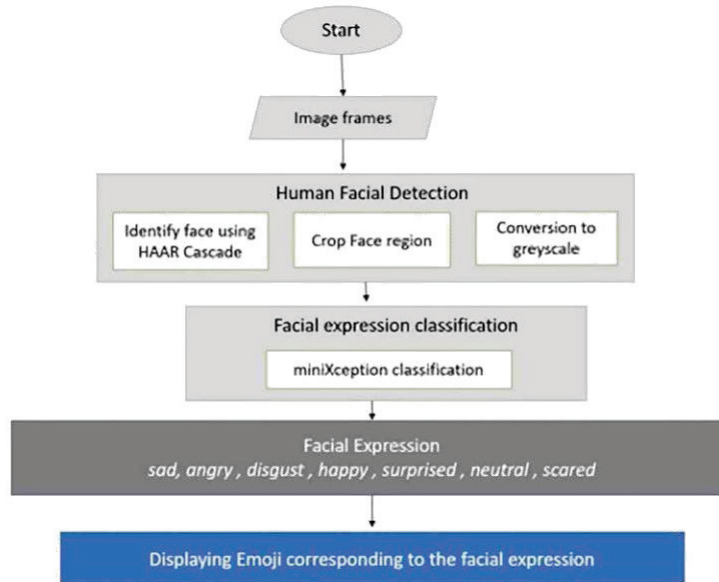


FIGURE 3. Proposed Architecture

The workflow starts with acquiring the FER 2013 dataset and pre-processing the images by scaling them in the range of -1 and 1. Images are scaled to [0,1] by dividing it with 255. Subtracting the value with 0.5 and multiplication by 2 changes the range to [-1,1]. Image shape is configured to produce an array of (48,48,1). After pre-processing, augmentations techniques are used to produce transformed images. The variations of images are generated at each epoch. ImageDataGenerator class is used to return the images that are transformed. Images are being loaded in batches which will reduce the memory usage. Images are randomly rotated at a degree of 10, followed by vertical and horizontal shifting by a pixel value of 0.1. Images are horizontally flipped which indicates that they are flipped over a horizontal axis. These techniques are stored as arguments in the class of ImageDataGenerator in a variable and later utilized while training the model. The proposed model consists of four residual modules where regularization (l2 regularization is in the argument of the separable convolutional layer) is done to discipline peaky weights and to check if every input is being considered. After regularization, batch normalization is applied to maintain mean activation closer to 0 and activation standard deviation closer to 1. RELU activation function is used to get the output of the node, which produces the output as 0 if negative values are present and the same output as the input if the value is equal or greater than 0. Max pooling with pool_size of (3,3) is carried out after another round of regularization and batch normalization. All these layers are added at the end of each residual network. Global average pooling is imposed to produce a vector which is obtained from averaging every element in the feature map in the final layer. Soft max activation function is used, which takes the vector value produced as input to obtain the normalized output between 0 and 1. This final output is passed as an argument while defining the variable of the proposed model.

Callback functions such as EarlyStopping is utilized to stop training if there is no improvement in the performance measure. Validation loss is the performance measure for ReduceLROnPlateau callback function, and its improvement is considered for modifying learning rate. Using fit generator, the proposed model is trained. Performance metrics are generated for further comparison. Using test dataset obtained from splitting of the FER 2013 dataset, testing is done. Real-time testing of the model begins with loading of the trained model. A function is defined to open the web camera and the face of the user gets detected using HAAR cascade as shown in Fig. 3. The prediction of the emotion from the user's facial expression is done using the trained Mini-Xception model. Classification is done and the emoji representing the emotion of the face is predicted. Each emoji represents one of the emotions among happy, sad, anger, fear, neutral, surprise and disgust.

RESULT ANALYSIS AND DISCUSSION

The dataset used in the training of the model is obtained from FER 2013 which consists of 30,000 images for the 7 expressions used such as angry, happy, sad, neutral, disgust, surprise and fear. 80% among this dataset is useful in training set and 20% in testing. Using testing set, the proposed model is tested. The performance of the proposed model is validated in real-time with the help of web camera by capturing the face images of the user. The input given by the user can be from the testing dataset (Feature present in the user interface) or can be dynamically captured from web camera. Different expressions are given as input to analyze the expression and to check whether the correct output is displayed or not.

When emotions are received, a sequential image is displayed to the right of the screen. It takes about 5 seconds to change from one emoji to another when the expression is changed. The accuracy obtained for training was 56% which is on par with many models which were created earlier. Testing accuracy of the proposed model is 55%. When a graph is plotted as shown in Fig. 4 between Accuracy and the number of epochs, it is observed that the accuracy increased on increasing the number of epochs and hence, the graph obtained is an increasing curve.

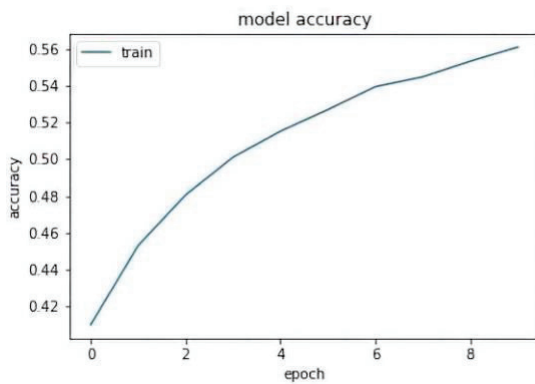


FIGURE 4. Graph for accuracy vs epoch for the proposed model

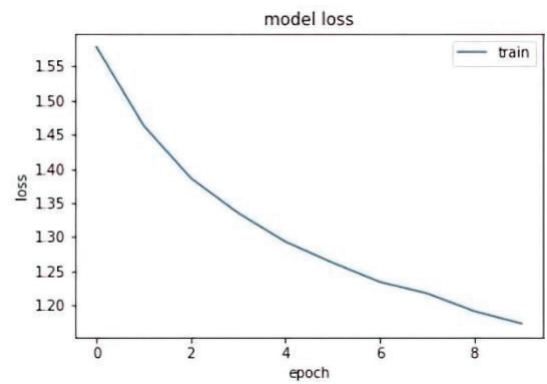


FIGURE 5. Graph for loss vs epoch for the proposed model

Similarly, when a graph is plotted between loss and number of epochs as shown in Fig. 5, results in decreasing curve; that means the loss function of the model decreases with the increase in number of epochs. Another model (Deep Convolutional Neural Network) is trained for comparison of performance. This time ELU (Exponential Linear Unit) activation function is used instead of RELU. ELU activation function can take negative.

values and has lower computational complexity but takes more time to train. Six convolutional layers are taken instead of depth wise separable layers. Dense layer is added to take input which are outputs of convolutional layers. Global average pooling is used in Mini-Xception which eliminates the need for fully connected layer. The output layer uses softmax activation to produce the output.

Figure 6 shows the comparison of accuracy of Mini-Xception and Deep Convolutional Neural Network. Epochs are taken on x-axis and Training accuracy is taken in y-axis. For 10 epochs the accuracy of Mini-Xception is 56% and the accuracy of deep convolutional neural network is 45%. Utilization of depth-wise separable convolutions over normal convolutions increases efficiency of the model.

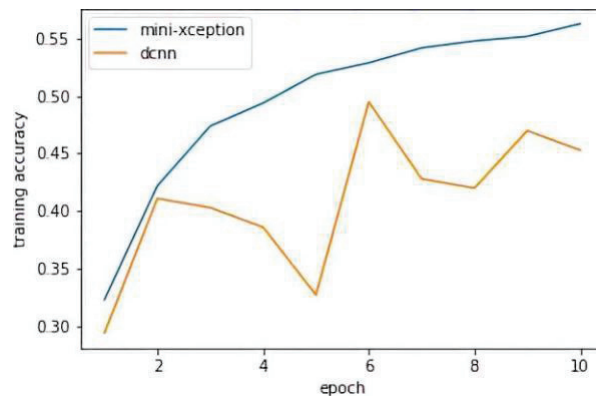


FIGURE 6. Accuracy comparison of Mini-Xception and deep convolutional neural network

FUTURE SCOPE AND APPLICATIONS

Facial expression detection is very useful to analyze the mood of the person and draw conclusions from that. Facial expression detection can be used for study of ASD (autism spectrum disorder), people with ASD can have difficulty in expressing themselves or convey their facial expressions. Facial expression recognition can be used as part of several available in-car systems which are being trained by machine learning algorithms. Adaptation of this application can help recognize if the driver of the car is falling asleep while driving. Facial expression detection can also be utilized by companies for their product marketing by making note of consumers' expressions while doing focus groups. Intelligent video analysis can be done while interacting in a teleconference where a video camera is used by both individuals to interact. Facial expressions detection techniques can be used by law enforcement in interrogation procedures for lie detection of the criminals. Using the video camera footage, features of the face can be extracted which are essential to classify innocent and guilty participants in the applications for security and surveillance. The facial expressions of the user can be detected and analyzed to know their state of mind. Products can be designed and employed according to the consumer's state of mind. Hence, facial expressions are essential for emotional analysis and detection which helps us to create viable and useful models.

CONCLUSION

Facial expression detection in real time can be a challenging task but highly advantageous. The expression of a person is valuable, which was concluded by the evidence that we have seen. The future scope for facial expression detection is limitless and with right expansion, can benefit us. The user's face is detected using a web camera. HAAR Cascade will extract the features of the face detected. This framework has been implemented with the Mini-Xception deep network algorithm which is efficient compared to other CNN algorithms as it takes less parameters which makes it less complex for further classification of extracted features. The application created has been tested with seven emotions (sad, angry, disgust, surprise, happy, neutral, scared) and an accuracy of 56% was produced which is on par with many models that were previously created. The model may perform less accurately than the user's expectations but can ensure facial expression detection and corresponding display of the emoji which furthers the seamless communication. So, as an

extension to this project, training should be done with different kinds of dataset; so that the model can learn in different aspects and make predictions in a more accurate way.

REFERENCES

1. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition-2001, IEEE Conference proceedings* (IEEE, Kauai, USA, 2001).
2. Küntzler, Theresa, Tim T. Höfling, and Georg W. Alpers, "Automatic Facial Expression Recognition in Standardized and Non-Standardized Emotional Expressions," in *Frontiers in Psychology*, Switzerland, (2021).
3. K.Revanth Reddy, N. Swapna Goud and G.Alekhya, in *International Journal of Trend in Scientific Research and Development*, (2019), pp. 1330-1333.
4. Ankur Ankit, Dhananjay Narayan and Alok Kumar, in *International Journal of Engineering and Advanced Technology*, (2019), pp. 256-258.
5. Nethra Chandrasekaran Sashikar – necsashi, Prashanth Kumar Murali – prmurali and Robert J Henderson – rojahend, "FEBEI- Face Expression Based Emoticon Identification CS - B657 Computer Vision," in *Association for the Advancement of Artificial Intelligence*, (2015).
6. M. Shamim Hossain, and Ghulan Muhammad, in *IEEE Access*, (2017), pp. 2281-2287.
7. Abita Devi, Kantesh kumar and Heena Gupta, in *international journal of Engineering and Development and Research*, (2017), pp. 326-329.
8. Daniel Llatas and Spiers, "Facial emotion detection using deep learning," *Uppasala University*, 2016.
9. Ninad Mehendale, "Facial emotion recognition using convolutional Neural Networks," *SN Appl. Sci.* 2, (2020).
10. Chandra Bhushan Singh, Babu Sarkar and Pushpendra Yada, in *Social Science Research Network, Rochester, NY*, (2021).
11. MD.Zia Uddin, Weria Khaksar and Jim Torresen," in *IEEE Access*, (2017), pp. 114-125.
12. Yingli Tian, Jeffrey F Cohn and Takeo Kanade," in *ResearchGate, Berlin, Germany*, (2011), pp. 487-517.
13. Kishore Babu, D., Ramadevi, Y., & Ramana, in *Arabian Journal for Science and Engineering*, pp. 705-714.
14. Babu, D. K., Ramadevi, Y., & Ramana, "PGNBC: Pearson Gaussian Naïve Bayes classifier for data stream classification with recurring concept drift," in *Intelligent Data Analysis*, pp. 1173-1191.
15. P. S. Prasad and V. Kumar Gunjan, "Feature Descriptors for Face Recognition," in *IEEE 17th India Council International Conference*, (IEEE,Kauai, USA, 2020).
16. Hoang Minh Phuong, Le Dung, T. de Souza-Daw, Nguyen Tien Dzung and Thang Manh Hoang, "Extraction of human facial features based on Haar feature with Adaboost and image recognition techniques," in *International Conference on Communications and Electronics*, (2012), pp. 302-305.
17. Kumar, S., Ansari, M. D., Gunjan, V. K., & Solanki, "On classification of BMD images using machine learning (ANN) algorithm," in *Springer*, (2019), pp. 1590-1599.
18. J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Imaging Systems and Techniques, IEEE Conference Proceedings*, (IEEE, Kauai, USA, 2015), (2015) pp. 1-6.
19. S. Minaee and A. Abdolrashidi," Deep emotion: Facial expression recognition using Attentional Convolutional network," in *Computer Vision and Pattern Recognition arXiv:1902.01019 [cs.CV]*, (2019).
20. Md. Jashim Uddin, Dr. Paresh Chandra Barman, Khandaker Takdir Ahmed, S.M. Abdur Rahim, Abu Rumman Refat and Md Abdullah-Al-Imran6," in *IOSR Journal of Electronics and Communication Engineering*, (2020), pp. 37-46.