IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 May 2022, accepted 30 June 2022, date of publication 5 July 2022, date of current version 11 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188730

RESEARCH ARTICLE

Efficient-SwishNet Based System for Facial Emotion Recognition

TARIM DAR[1], ALI JAVED[1], (Member, IEEE), SAMI BOUROUIS[2], HANY S. HUSSEIN[3,4], (Senior Member, IEEE), AND HAMMAM ALSHAZLY[5]

[1]Department of Software Engineering, University of Engineering and Technology, Taxila, Taxila 47050, Pakistan
[2]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[3]Electrical Engineering Department, College of Engineering, King Khalid University, Abha 62529, Saudi Arabia
[4]Electrical Engineering Department, Faculty of Engineering, Aswan University, Aswan 81528, Egypt
[5]Faculty of Computers and Information, South Valley University, Qena 83523, Egypt

Corresponding author: Ali Javed (ali.javed@uettaxila.edu.pk)

This work was supported in part by the Deanship of Scientific Research at King Khalid University through Research Groups Project under Grant RGP.2/16/43; in part by the Taif University Researchers Supporting Project under Award TURSP-2020/26; and in part by the Multimedia Signal Processing (MSP) Research Laboratory, University of Engineering and Technology, Taxila, Pakistan.

ABSTRACT Facial emotion recognition (FER) is an important research area in artificial intelligence (AI) and has many applications i.e., face authentication systems, e-learning, entertainment, deepfakes detection, etc. FER is still a challenging task due to more intra-class variations of emotions. Although prior deep learning methods have achieved good performance for FER. However, still there exists a need to develop efficient and effective FER systems robust to certain conditions i.e., variations in illumination, face angles, gender, race, background settings, and people belonging to diverse geographical regions. Moreover, a generalized model for the classification of human emotions is required to be implemented in computer systems so that they can interact with humans according to their emotions and improve their interaction. This work presents a novel light-weight Efficient-SwishNet model for emotion recognition that is robust towards the aforementioned conditions. We have introduced a low-cost, smooth unbounded above and bounded below Swish activation function in our model. Property of unboundedness helps to avoid saturation while smoothing helps in optimization and generalization of the model. Performance of the proposed model is evaluated on five diverse datasets including CK+, JAFFE, FER-2013, KDEF, and FERG datasets. We also performed a cross-corpora evaluation to show the generalizability of our model. The proposed model achieves a very high recognition rate for all datasets that prove the merit of the proposed framework for both the human facial images and stylized cartoon characters. Moreover, we conducted an ablation study with different variants of our model to prove its efficiency and effectiveness for emotions identification.

INDEX TERMS EfficientNet, efficient-SwishNet, facial emotion recognition, human computer interaction (HCI), Swish activation.

I. INTRODUCTION

Emotions are an obvious element of communication which are exhibited in several forms that may or may not be interpreted by the human eye [1]. Facial expression recognition (FER) is an essential research area in the field of AI. FER is defined as a technique of defining the mental state of humans by evaluating the motion or position of facial components such as blinking of eyes, expansion or contraction of eyelids, movement of lips, skewed eyebrows, and creasing nose. The expression on the face relies on skeptical things like the movement of facial muscles, surroundings, and mind condition [2]. Human facial expression is generally divided into seven classes i.e., angry, disgust, fear, happiness, surprise, sadness, and neutral [3]. Researchers are trying to impart intelligence in computer systems by recognizing the emotions of humans which helps to improve the interaction between the computer and humans. FER has many

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

applications in various domains i.e., face authentication systems, e-learning, detecting emotions of drivers while driving, systems helping the disabled person, healthcare, entertainment, deepfakes detection, etc., [3]. Facial emotion analysis will also help in obtaining customer feedback for services or products as most of the customers do not give feedback during the surveys. Companies can increase the quality of products and analyze customer satisfaction by demonstrating emotional expression [4]. Moreover, the significance of facial emotion analysis is also being investigated recently for deepfakes detection [35] where artificial abnormalities in the emotions are examined to identify whether the video is real or fake. So, emotions are an integral part of human lives and play a vital role in how humans perceive or understand things [5].

Human and computer interaction (HCI) plays a vital role in regular activities nowadays. We are moving towards the generation where most activities are carried out via PC instead of written documents. Facial expression recognition is also important in non-verbal communication that can be taught to machines or robots so that they can interact with users according to their emotions. Understanding the variations in human emotions are still challenging for machines. From the last few decades, many manual and automated methods were developed to assist the emotion analysis process. But still, an efficient and more accurate model is required to better address this problem.

From the last few years, research on human emotion recognition has helped to understand the psychological behavior of humans. Mehendale *et al.* [12] presented a two-part model named FERC; the first part of the model removed the background, and the second part extracted the facial expression vector (EV). FERC was tested on multiple datasets and attained 96% accuracy on the NIST dataset. This approach [12] worked well for datasets having images in the range of 2000 to 10000. However, the performance of this method was degraded when evaluated with the dataset having images fewer than 2000 or more than 10000. Moreover, the appearance of multiple faces in images at an equal distance from the camera leads to the failure of the algorithm. Face orientation and shadow on the face also affects the performance of this model [12] as all values of the feature vector are not obtained due to shadow or face orientation. Ramdhani *et al.* [4] proposed an expression identification system based on convolutional neural networks (CNN) and compared two CNN configurations of batch sizes 8 and 128. The purpose of this approach [4] was to gather customer satisfaction with the product. This approach was validated on FER-2013 and self-created datasets against the four emotions.

Many existing studies of facial expression examination are evaluated on small or posed datasets against six emotions. These approaches are not robust and consume more resources if deployed in real-world systems. They also fail to perform well in case of the presence of accessories like glasses, hair, etc., on the face or different physical conditions like illumination or face angle condition. The recognition rate is also low due to differences in appearance and characteristics of people belonging to different geographical regions. So, an efficient and robust model is needed which uses fewer resources and can be deployed in any embedded system without performance degradation. Moreover, previous research works are not subjected to cross corpora evaluation to build a generalized model with less overfitting.

To overcome the limitations i.e., high computational complexity, non-generalized model, and low recognition rate in the absence of frontal face images or images with faces oriented at different angles, we proposed an Efficient-SwishNet deep learning (DL) model to automatically detect seven facial emotions using the face images of diverse datasets. The proposed FER model is able to detect diverse facial emotions more accurately and computationally efficient as compared to the existing models. The proposed model employs a customized EfficientNet-b0 architecture along with the addition of Swish activation function for the classification of different facial emotions. EfficientNet is a CNN-based architecture and scaling model which utilizes a compound coefficient to scale the network's width, depth, and resolution. Moreover, the EfficientNet-b0 model has the least number of parameters and floating-point operations (FLOPS) than other CNN's architecture [36]. The Swish activation function is a simple multiplication function of input value with a sigmoid function. This function does not instantly change the direction on negative values rather it smoothly changes the direction and then moves upwards. The Swish activation function allows a small range of negative values, which facilitates to capture the patterns in data. A smooth curve is obtained as a result of this activation function which aids to optimize the model in terms of convergence with minimum loss. Moreover, employing the Swish activation in EfficientNet-b0 architecture makes the model optimal and computationally efficient. The main contributions of this paper are as follows:

- We propose a lightweight and effective Efficient-SwitchNet DL model with a Swish activation function to automatically detect eight diverse facial emotions more accurately.
- The proposed model is robust to variations in illumination conditions, face angles, gender, race, background settings, and morphological appearance of people belonging to diverse geographical regions.
- We also performed extensive experiments on multiple diverse datasets containing the facial emotions of real and stylized cartoon characters' images and including the cross corpora evaluation to demonstrate the generalization ability of the proposed model for FER.

The rest of the research paper is organized as follows: Section II reviews the literature work. In Section III, the proposed methodology is discussed in detail. The details of each dataset are provided in Section IV. Section V presents the experimental results, analysis, and discussion. And conclusions are drawn in Section VI.

II. LITERATURE REVIEW

Facial expressions as an indication of feelings and emotional information are mostly used for the identification of emotions. Emotion recognition approaches are categorized into three types: deep learning methods, appearance-based, and geometric-based feature extraction approaches.

A. APPEARANCE AND GEOMETRIC BASED FEATURE EXTRACTION APPROACHES

Early research was mostly focused on the approaches of geometric and appearance-based feature extraction methods. Appearance-based features represent the texture of face while the geometric-based features represent the shape of the face [6]. Appearance features are effective in finding fine color and texture changes in the face while geometric features are good in classifying the facial expressions [7]. Appearance-based feature extraction approaches are more popular than geometric-based approaches due to their computational efficiency over the latter. Spatial filters are either applied to the entire face or to a selected area of the face in appearance-based approaches. Two important and popular techniques of appearance-based methods are Gabor Wavelet and Local Binary Patterns (LBP). Different variants of LBPs are used in past research [17]. The algorithm presented in [16] was based on the fused features extracted from the LBP, oriented FAST and Rotated BRIEF (ORB) descriptors. After this, SVM was used with a 10-fold cross-validation approach to identify the emotions. This model [16] obtained an accuracy of 92.4%, 99.2%, and 84.2% for subject-dependent scenario while 88.4%, 93.2%, and 79.8% for subject-independent scenario on Japanese Female Facial Expression (JAFFE), Extended Cohn-Kanade (CK+), and MMI datasets, respectively. Kola *et al.* [17] proposed an LBP-based approach by introducing the concept of an adaptive window of different sizes for feature extraction. This method was evaluated on Cohn-Kanade (CK) and JAFFE datasets against six and seven emotions. The recognition rate on the JAFFE dataset against 6 and 7 emotion classes was 88.3% and 92.9%. On the CK dataset, this method achieved an accuracy of 96% for 6 emotional classes and attained 93.9% accuracy when trained and tested for 7 classes of emotions. Moreover, this model [17] was evaluated on the Facial Expression Research Group (FERG) dataset with an accuracy of 96.7% and FEI dataset with an accuracy of 98.9%. Jiang *et al.* [37] presented a Gabor convolutional network (GCN) network for FER. This model was tested with varying numbers of layers (depth) and units per convolutional layer (width) and compared it against the well-known CNN architectures. This model [37] was evaluated on three benchmark datasets namely Facial Emotion Recognition (FER-2013), FERPlus, and RAF with an accuracy of 72.54%, 87.75%, and 86.77%, respectively on these datasets.

Previous works have also employed the geometric features extraction-based approaches for emotion classification. Ounachad *et al.* [18] presented a geometrical method based on hamming distance with face ratio for emotion recognition. The effectiveness of this model [18] was checked on the Warsaw set of emotional facial expression pictures dataset and achieved a recognition rate of more than 93%. An automated system was presented in [34] to measure the movement of facial muscles and identification of emotions. This system obtained geometric features of face and estimated displacement of feature points between the expression and neutral frames and then applied a two-stage fuzzy reasoning model in which the first stage was designed for classification of Action Units (AUs) from geometric measurement vector. Whereas the second stage was intended for the classification of five basic emotions i.e., anger, surprise, sadness, fear, and happy from the AUs. The model was tested on CK+ dataset and acquired an accuracy of 90%. Furthermore, Murugappan *et al.* [24] developed a triangulation method for extracting the geometric features. Firstly, Haar-like features were used to detect the subject's face. Next, eight computer-generated markers referred as AUs were placed on the defined location of the face, and a combination of three AUs was used to make five triangles. The Inscribed circle circumference, the Inscribed circle area of a triangle (ICAT), and area of the triangle were extracted as features to classify the facial emotions. This method was used for the classification of six emotions (fear, sadness, anger, disgust, surprise, and happy) using six types of machine learning (ML) classifiers on a custom dataset containing 85 subjects (55 males and 30 females). The ICAT feature with Random Forest provided the maximum accuracy of 98.17%.

B. DEEP LEARNING MODELS

Modern researchers use deep convolutional neural networks for combining feature extraction and emotion classification into one process [7]. CNN was specially designed for image classification tasks. Talegaonkar *et al.* [8] presented a DL model based on CNN for real-time emotion classification using the webcam. The model was designed to recognize the emotions of the user while watching movie trailers or listening to the video lectures. This model [8] was evaluated on FER-2013 dataset and achieved 60.12% accuracy. This model gives poor performance on the recognition of fear and disgust emotions. Ozdemir *et al.* [5] presented a LeNet architecture-based DL system for the classification of seven emotions. Performance was evaluated on the dataset created by merging the Karolinska Directed Emotional Faces (KDEF), JAFFE, and their custom dataset, and achieved an accuracy of 91.81%. Although this system provides better accuracy than [8], however, this system is unable to perform well on the prediction of sad emotion.

Developing a model that is efficient both in terms of time and space is still a difficult task. MicroEXPNet [9] is a small and fast CNN model designed for face authentication from frontal face images in mobile phones. This model [9] was trained using the InceptionResnet-v3 followed by the application of knowledge distillation. The accuracy of this model fluctuates by changing the temperature parameters. Minaee *et al.* [1] presented a network based on attentional convolutional networks for classification of emotions from the face images. The model [1] paid attention to special regions on the face to detect different emotions. Performance of this model [1] was evaluated on four different datasets including CK+, JAFFE, FER-2013, and FERG. The highest accuracy of 98% was achieved on CK+ dataset whereas, lowest accuracy of 70.02% was obtained on the FER-2013 dataset due to imbalanced intra-class variation among different classes in the dataset.

Some research works [10-11] combined the emotion detection task with gender recognition. Arriaga *et al.* [10] developed two CNN models for emotion and gender recognition in real-time. First model called fully sequential CNN eliminated the fully connected layers completely. Second model called Mini-Xception also eliminated fully connected layers and added depth-wise separating convolutional and residual modules. Depth-wise separation of convoluted layers helps to reduce the computation. A real-time system was implemented for validation of the model [10] but this system is biased for western face images and person with glasses is mostly classified as angry and male. Moreover, the usage of accessories like glasses on faces affects the gender and emotion classification process. Yar *et al.* [11] proposed two CNN models for face emotion detection and gender recognition in real-time videos. The performance of this model was evaluated on FER-2013 and Internet Movie Database (IMDB) gender datasets and attained an accuracy of 90%.

For automatic detection of facial expression, a shallow convolution neural network (SCNN) was presented in [13]. The model consisted of two convolution layers followed by a fully connected layer. This model [13] was evaluated on two datasets i.e., CK+ and JAFFE, with an accuracy of 99.49% and 93.02%, respectively. The model has a low confidence score for happy emotion in JAFFE because happy emotion images contain similar features as other emotions like anger and disgust [13]. Similarly, Dubey *et al.* [14] demonstrated a DL framework based on transfer learning using the VGG16 architecture. The model [14] removed the top layers of VGG16 and added new layers such as Flatten, Drop, Dense, and dense-SoftMax layers for the classification of emotions. The model attained an accuracy of 93.75% on the JAFFE dataset and 94.84% on the CK+ dataset. Zhong *et al.* [15] proposed an efficient and simplified model named SE-SResNet18 based on Residual Network (ResNet18) and Squeeze-and-Excitation (SENet). The performance of SE-SResNet18 [15] was checked on CK+ and FER-2013 datasets and achieved an accuracy of 74.143% with resized images of 44×44 pixels on FER-2013. Accuracy of 95.253% was achieved on the CK+ dataset with 10-fold validation. It was also shown that deeper networks hardly improve the recognition accuracy.

Chowdary *et al.* [38] used four pre-trained deep CNN models i.e., ResNet50, Inception-v3, VGG19, and MobileNets for the emotion classification task. Performance was evaluated on 918 images of CK+ dataset and attained an accuracy of 98.5% on MobileNet50, 97.7% on ResNet50, 94.2% on Inception-v3, and 96% on the VGG19 model. Pranav *et al.* [39] developed a two-layer convolution network to classify five different emotions i.e., angry, sad, surprise, neutral and happy. This model was tested on a custom dataset of 2550 images with an accuracy of 78.04%.

III. PROPOSED METHODOLOGY

In order to solve the above-stated problems, a novel Efficient-SwishNet deep neural network is presented for the classification of different human facial emotions from multiple and diverse datasets. The details of our DL model are explained in the subsequent sections.

A. PREPROCESSING

In the pre-processing step, all the images of each dataset are resized to 224 × 224 with three channels. This resolution for the input image is the requirement of our model. After preprocessing, images are fed to our customized Efficient-SwishNet model to extract the reliable features and later classify the emotions of seven different categories.

B. ARCHITECTURE DETAILS

In previous CNN models such as ResNet50 or ResNet152, more layers were added to enhance the model performance for classification but at the expense of higher computational cost. Three different types of scaling i.e., depth, width, and resolution or a combination of any two were performed in earlier CNNs to boost the classification performance of the model. Depth-wise scaling means an addition of more layers which helps in capturing complex features and increases the generalizability of the model while width-wise scaling indicates a total number of convolution channels involved in the convolution layer which facilitates in capturing fine-grained features. Similarly, images of higher resolution are used in complex tasks as they contain more information. Instead of scaling one or two dimensions, EfficientNet uses compound scaling of three parameters of width, depth, and resolution.

EfficientNet-b0 was developed by multi-objective neural architecture search that helps to improve the accuracy as well as efficiency or FLOPS. EfficientNet model uses a compound scaling coefficient that uniformly scales three dimensions of network width, depth, and resolution. EfficientNet models have proved to be more accurate and efficient than the traditional CNN architecture and have also shown good performance at transfer learning [36]. The number of parameters and FLOPS of EfficientNet is smaller than those of previous CNN architectures. We utilize EfficientNet-b0 from the EfficientNet family (i.e., EfficientNet b0 to b7) for features extraction and transfer learning because it has the fewest parameters and FLOPS on ImageNet, shown in an experiment in [36]. The most accurate model in EfficientNet family is the EfficientNet-b7, however, it is computationally expensive having the highest FLOPS and parameters. So, we chose EfficientNet-b0's architecture for customization since it is more cost-effective and efficient than other EfficientNet models. The aim of using the EfficientNet-b0

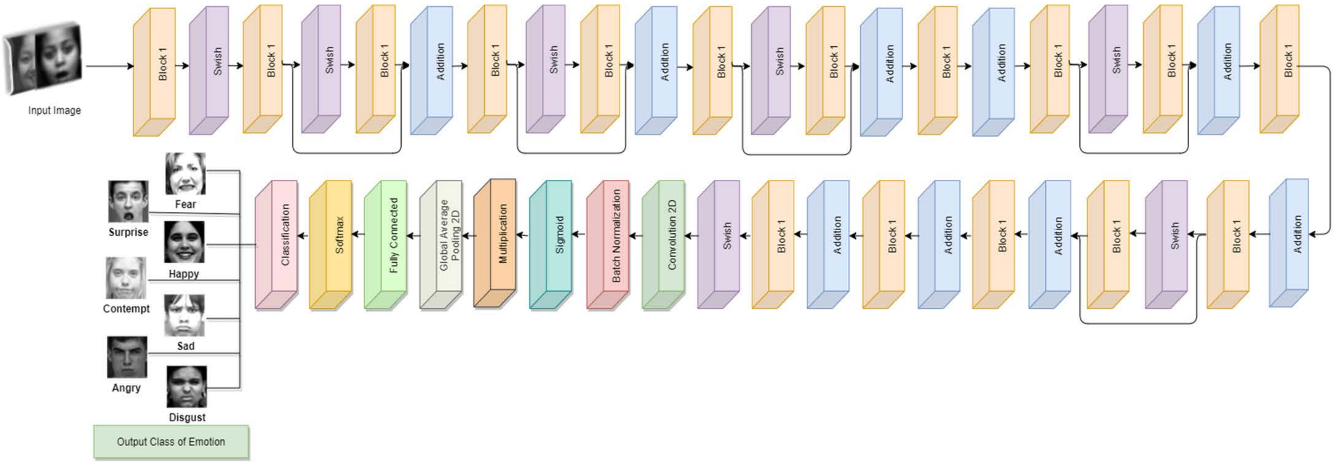


FIGURE 1. Architecture diagram.

model for customization is to increase the accuracy while keeping it computationally efficient. EfficientNet is the cost-efficient, robust and up-to-date model. In addition to squeeze-and-excitation blocks, EfficientNet-b0 architecture is based on the inverted bottleneck residual blocks of MobileNet-v2 [41]. The input size for EfficientNet-b0 is 224 × 224×3. The family of EfficientNet consists of eight models ranging from b0 to b7. The b7 model has the highest number of parameters but it has a more computational load. As EfficientNet-b0 has the lowest complexity and greater speed, therefore, we customized the EfficientNet-b0 and employed it in this study for the efficient classification of facial emotions.

1) ARCHITECTURE OF EFFICIENT-SWISHNET

Figure 1 shows the overall architecture of our Efficient-SwishNet model containing 297 layers in total. Our model contains both the point-wise and depth-wise convolution layers. Depth-wise convolution is performed on each channel of the input independently while point-wise convolution uses a 1 × 1 kernel size that passes through every point of input. After splitting the data into train and test sets and preprocessing, images are fed to the proposed model for feature extraction from the input image. The last layer performed the global average pooling, calculating an average value of the feature map in 2 dimensions i.e., height and width, and pass single output to the fully connected layer. The last three layers (fully connected layer, softmax, and classification layer) are used for the classification of an image to detect the category of emotions.

The sequential details of the proposed model are shown in Table 1, which includes the layers' name, activation in these layers, and total learnable parameters of individual layer. Information of starting 17 layers (Block 1) and last 10 layers of our model is described in Table 1. Input image size is 224 × 224×3 and reduced to half i.e., 112 × 112 in next convolution 2D layer. Point-wise and Depth-wise convolution layers are presented to extract fine details of the features. Layers up to the global average pooling extract the features and the remaining three layers are used to perform the classification task.

2) ARCHITECTURAL DETAILS OF BLOCK 1

Figure 2 provides insight of Block 1 (Figure 1) used in the proposed model, which is the main part of our architecture consisting of 17 layers. Convolution layers are the main block of this model like other CNNs and have kernel sizes of 3×3 or 5×5. Filters of large size improve the model accuracy and efficiency and capture high-resolution patterns while small size filters capture the low-resolution patterns. Layers after convolution layers are scaled down to reduce the feature map size but the width is scaled up which facilitates in improving the accuracy. The number of filters in each convolution layer is increased and reached 280 till the fully connected layer. The output of the convolution layer in the form of features map is passed to the next layers for further processing.

3) SWISH ACTIVATION FUNCTION

To improve the performance and get better classification results, we introduced the Swish activation function in our customized model after every convolution 2D followed by the batch normalization layer. Swish is a non-monotonic, smooth unbounded above and bounded below activation function. Unboundedness helps to avoid saturation, and strong regularization effects are obtained from the bounded below property. Smoothing helps in generalization and optimization, while non-monotonicity improves the gradient flow and provides robustness to different learning rates. Swish is a simple function and experiments show that it performs rather well than the most popular ReLU function on complex domains of image classification and machine translation [28]. Moreover, Swish passes the small number of negative weights while ReLU assigns zero to negative weights. The Swish function is computed as:

$$f(x) = x \times \text{sigmoid}(\beta x) \quad (1)$$

TABLE 1. Summary of starting block1 of proposed architecture along with the last ten layers.

Layer	Activation	Learnable parameters
Input Layer	224×224×3	0
Convolution2D	112×112×32	896
Batch Normalization	112×112×32	64
sigmoid	112×112×32	0
Element Wise Multiplication_1	112×112×32	0
Grouped Convolution	112×112×32	320
Batch Normalization	112×112×32	64
Sigmoid	112×112×32	0
Element Wise Multiplication_2	112×112×32	0
Global Average Pooling	1×1×32	0
Convolution	1×1×8	264
Sigmoid	1×1×8	0
Element Wise Multiplication	1×1×8	0
Convolution	1×1×32	288
Sigmoid	1×1×32	0
Element Wise Multiplication	112×112×32	0
Convolution	112×112×16	528
Batch Normalization	112×112×16	32
...	...	...
...	...	...
Batch Normalization	7×7×320	640
Swish	7×7×320	0
Convolution	7×7×1280	410880
Batch Normalization	7×7×1280	2560
Sigmoid	7×7×1280	0
Element Wise Multiplication	7×7×1280	0
Global Average Pooling	1×1×1280	0
Fully Connected	1×1×7	8967
Softmax	1×1×7	0
Classification	1×1×7	0

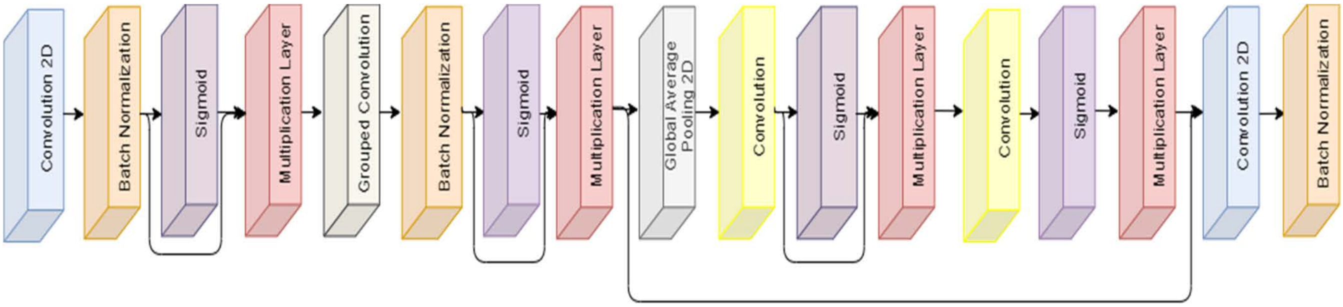


FIGURE 2. Details of block 1.

To make our model more efficient and select the best activation function which not only increases the accuracy of our model but is also computationally cost-effective, we conducted an experiment as described in *Section V-E2*. This experiment proves that addition of Swish activation function in the proposed model makes it computationally more efficient as Swish has less training time as compared to other activation functions. Also, an ablation study was conducted in *Section V-E*1 that shows that the proposed model with the Swish activation function achieved the best results.

IV. DATASETS

To evaluate the performance of our model, we have selected five different and diverse datasets i.e., JAFFE [32], CK+ [27], KDEF [25], FER-2013 [26] and FERG [19]. These datasets include posed as well as spontaneous images. Not only frontal images but images captured from different angles are also included in our experiments. Moreover, a stylized characters dataset is included to show the effectiveness of the proposed model. These datasets contain images of people belonging to different geographical areas of the world including Asian, European, American, and African. Details of each dataset used for experiments are mentioned in the subsequent subsections.

A. JAPANESE FEMALE FACIAL EXPRESSION DATASET

This dataset [32] consists of 213 gray posed images of 256×256 resolution of ten different Japanese female models. Each subject has 3 or 4 samples of seven facial emotions

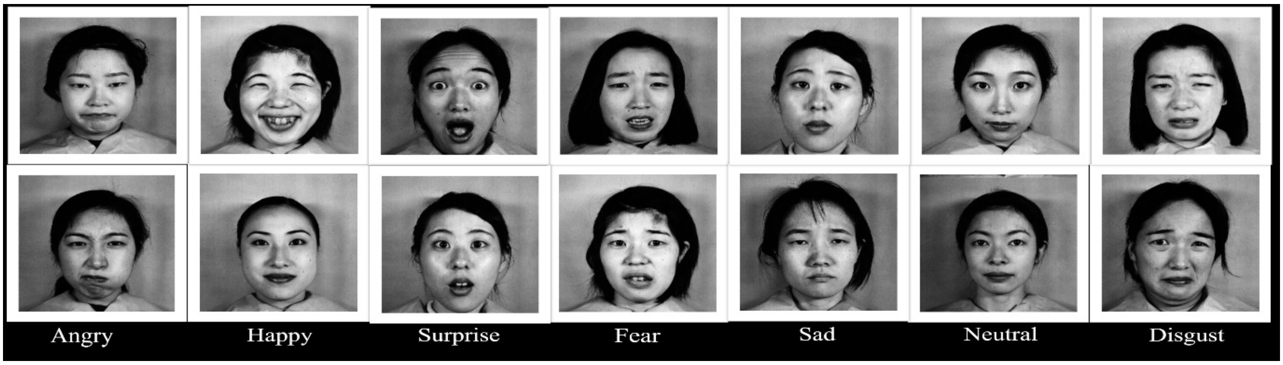


FIGURE 3. Few images of JAFFE dataset.

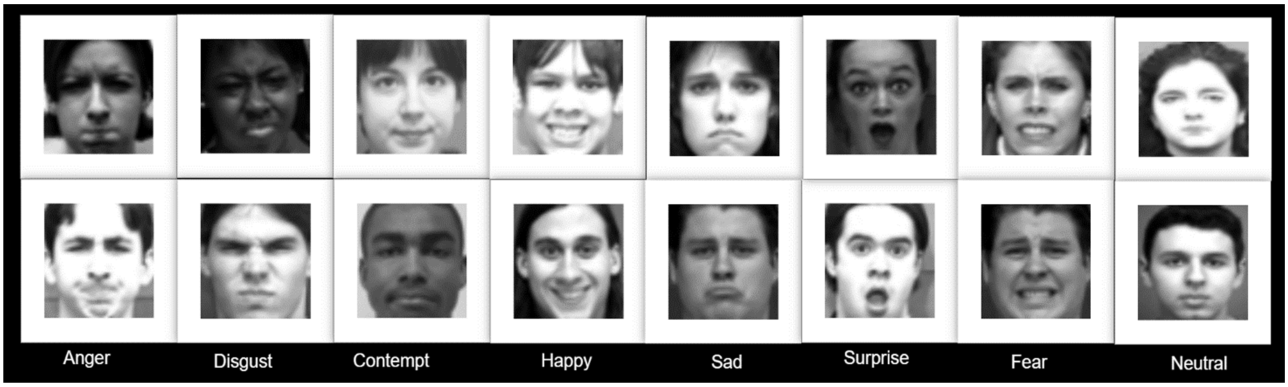


FIGURE 4. Few images of CK+ dataset.

(one neutral and six basic emotions). This lab-controlled dataset is biased in the context of gender (only female) and ethnicity (just Japanese). In our experiments, all 213 images are used (anger: 30; disgust: 29; fear: 32; neutral: 30; happy: 31; sad: 31; and surprise: 30 images). Some samples of each emotion from the JAFFE dataset are presented in Figure 3.

B. EXTENDED COHN-KANADE (CK+) DATASET

This dataset [27] is a publicly available dataset for action units and human expressions identification. It includes both the posed and non-posed (spontaneous) human emotions. CK+ contains 593 sequences from 123 subjects within a range of 18 to 30 years of age. Among them, females were 65%, 15% were African American and Asians or Latino were 3%. Each expression contains images from neutral to peak facial expression. We have considered the last three frames of each sequence. The first frame of each sequence is considered neutral. There is a total of 1200 images of anger, disgust, fear, happiness, sad, surprise, contempt, and neutral. Some images of this dataset are shown in Figure 4.

C. KAROLINSKA DIRECTED EMOTIONAL FACES DATASET

The Karolinska Directed Emotional Faces dataset contains a total of 4900 images of seven human facial emotions [25]. This dataset comprises artificially created emotion images of 70 individuals (35 males and 35 females). Images were taken from five different profile views of each individual including straight or frontal, full left, full right view, half left, and half right angles. This dataset is challenging as one eye and one ear from the complete face is observable in the case of full left and full right profile view, making the emotion recognition more complex. Figure 5 indicates few samples of the KDEF dataset.

D. FER-2013 DATASET

The dataset [26] contains grayscale images of 48 × 48 resolution. Images are labeled with six emotional categories of angry, fear, happy, sad, surprise, and disgust along with neutral. The database was introduced in ICML-2013 challenges and was created by Google image search API. FER-2013 is a more challenging and diverse dataset than other datasets available for emotion classification. FER-2013 contains the images of males, females, and children with face occlusion (face is covered with the hand, hairs, etc.), half-faced and low contrast images, and images with accessories like eyeglasses. It also contains non-face images or faces with text on them. Figure 6 represents few samples from the FER-2013 dataset.

E. FACIAL EXPRESSION RESEARCH GROUP 2D DATABASE

FERG database [19] consists of 2D images of six stylized characters (Aia, Bonnie, Mery, Ray, Malcolm, and Jules) with annotated facial expressions. The database contains 55767 annotated facial images of seven types of expressions

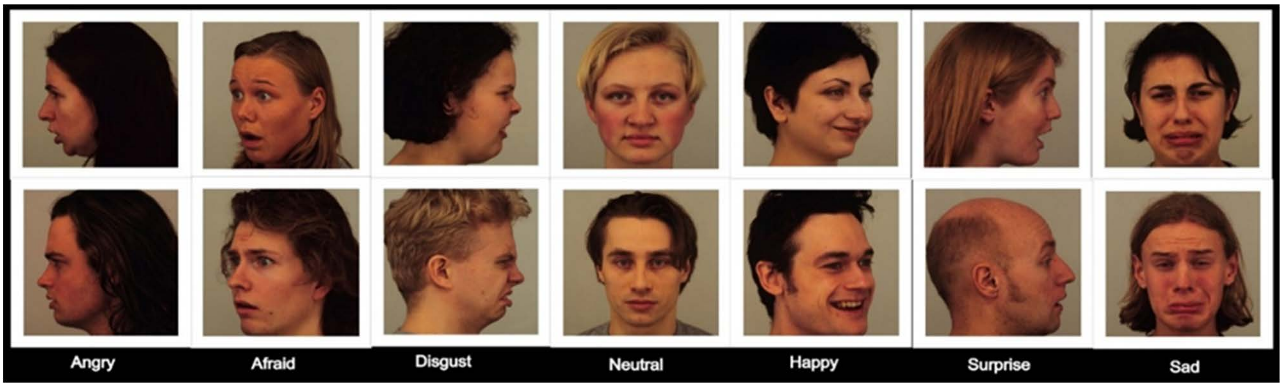


FIGURE 5. Sample images from KDEF dataset.

FIGURE 6. Some Images of FER-2013 dataset.

i.e., anger, disgust, fear, joy, neutral, sadness, and surprise. The characters were modeled using the MAYA software [42] and rendered out in 2D to create the images. Figure 7 shows a few sample images of the FERG database.

V. EXPERIMENTS AND RESULTS

This section provides the details of the experimental setup of the proposed model based on different facial expression datasets and a discussion of the results obtained by the proposed model. We also compared the performance of our model with the prior methods to justify the efficacy of the proposed model. Results of the ablation study and cross-corpora evaluation are also described in the subsections.

A. IMPLEMENTATION DETAILS

For all experiments after the preprocessing step, the model is trained for all datasets except JAFFE using the Adam optimizer and learning rate = 0.0004. Other parameters are: Batch size = 32, Epochs = 20 and Shuffle=every-epoch. JAFFE is a small dataset, so batch size is reduced to increase the number of images per epoch. We trained our model on JAFFE with Batch size = 16, Epochs = 80, other training parameters remain the same. The model is implemented using Matlab 2021a. For model implementation and execution, we used the machine with the following specifications: 8GB RAM, 1.6 GHz CPU, 930 GB Hard disk, and Windows10 Pro.

B. EVALUATION METRICS

To evaluate the performance of the proposed model, we calculated the accuracy, mean precision, and mean recall which are the standard metrics also used by the contemporary methods for facial emotion recognition.

1) ACCURACY

The most important metric in multi-class classification is accuracy. It is calculated by dividing the sum of true negatives (TN) and true positives (TP) instances of the emotion class by the overall number of instances in the emotion class. The accuracy of the classifier is computed as:

$$Overall\ Accuracy = \frac{\sum_{i=1}^{n} TP_i + TN_i}{TP + TN + FP + FN} \tag{2}$$

2) MEAN PRECISION

Precision represents the true positive predictions made by the proposed model in the case of each emotion class. Mean precision of the model in case of multi-class classification is

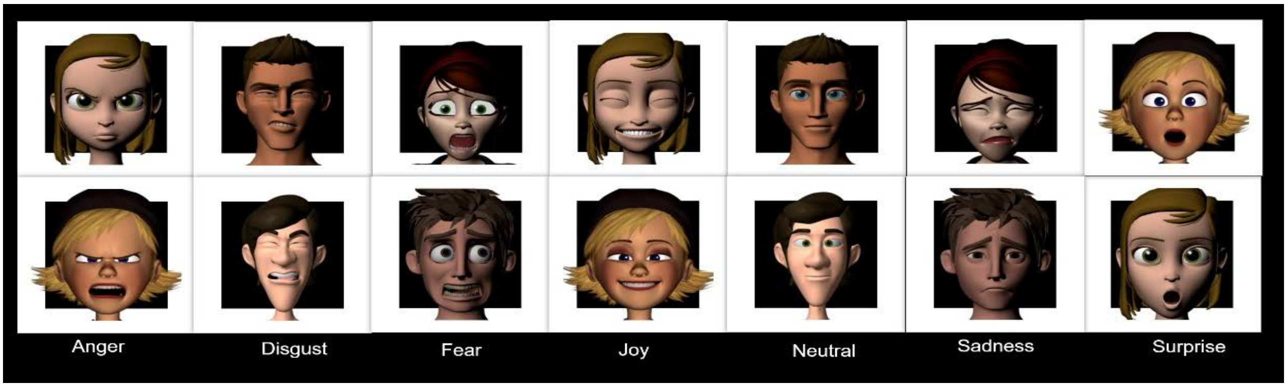


FIGURE 7. Sample images of FERG dataset.

calculated as:

$$\text{Mean } Precision = \frac{\sum_{i=1}^{n}(\mathrm{TP_i})}{\sum_{i=1}^{n}(\mathrm{TP_i}+\mathrm{FP_i})} \tag{3}$$

3) MEAN RECALL

Recall specifies the number of positive class predictions made out of all the positive instances in the emotion class present in the dataset. We computed the mean recall as follows:

$$MeanRecall = \frac{\sum_{i=1}^{n}(\mathrm{TP_i})}{\sum_{i=1}^{n}(\mathrm{TP_i}+\mathrm{FN_i})} \tag{4}$$

4) MEAN SPECIFICITY

It is defined as a fraction of negative emotion classes which are correctly classified as negative by the classifier. Specificity is also called true negative rate represented as TNR. We computed the mean specificity as follows:

$$MeanSpecificity = \frac{\sum_{i=1}^{n}(\mathrm{TN_i})}{\sum_{i=1}^{n}(\mathrm{TN_i}+\mathrm{FP_i})} \tag{5}$$

5) MACRO F-1 SCORE

Macro F-1 score is calculated by taking the unweighted arithmetic mean of F-1 score of each emotion class. Macro F-1 score in terms of TP, false positive (FP) and false negative (FN) is expressed as:

$$Macro\ F-1Score = \frac{\sum_{i=1}^{n}(\mathrm{TP_i})}{\sum_{i=1}^{n}(\mathrm{TP_i}+0.5(\mathrm{FP_i}+\mathrm{FN_i}))} \tag{6}$$

In Eqs. (2) - (6), n is the total number of classes, TP$_i$, TN$_i$, indicate the true positive and true negatives whereas FP$_i$, and FN$_i$ denote the number of false positives, and false negatives, respectively for the emotion class i.

C. PERFORMANCE EVALUATION ON DIFFERENT DATASETS

We designed multiple experiments to investigate the efficacy of the proposed method for facial emotion recognition on a variety of diverse datasets i.e., CK+, JAFFE, KDEF, FER-2013, and FERG. The details of all of these experiments are presented in the subsequent sections.

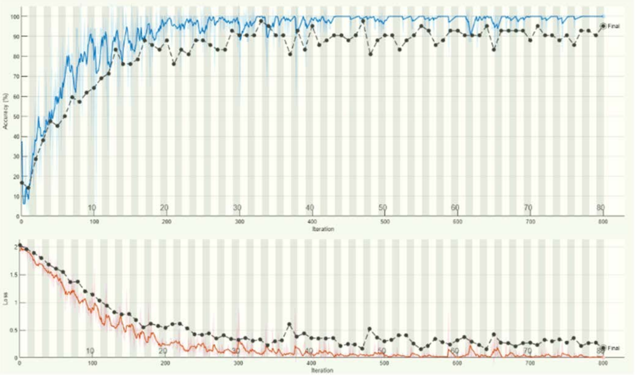


FIGURE 8. Progress of model in terms of accuracy and loss on JAFFE dataset.

1) RESULTS ANALYSIS ON JAFFE DATASET

To check the performance of FER on a small posed dataset, we evaluated our method on the JAFFE dataset. We conducted an experiment to distinguish different classes of emotions. To avoid overfitting on the small JAFFE dataset, we employ the data augmentation technique as the original dataset contains only 213 images. Options used for dataset augmentation include the following operations i.e., rotation, reflection, shear, etc. JAFFE dataset is split into train (80%) and test set (20%). After the training, model is tested on the test set to evaluate the model performance. The model achieved an overall accuracy of 95.2%, mean precision of 95.9%, and mean recall of 95.2% on this dataset although some of the images in this dataset are mislabeled. These results, with the accuracy, precision, and recall all greater than 95%, on a dataset having mislabeled class problem shows the efficacy of our method for FER. Figure 8 depicts the performance of our model on JAFFE dataset in the form of accuracy and loss graph.

We have also designed an experiment to generate the confusion matrix of our method on JAFFE dataset. Confusion matrix is used to present experimental results for each dataset as it gives visual representation of false positive rate (FPR) and true positive rate (TPR). Confusion matrix analysis is designed to better analyze the false acceptance and rejection

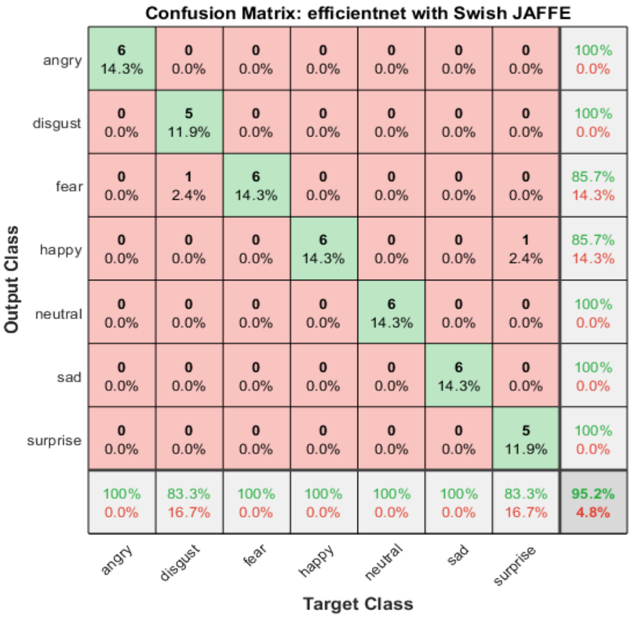


FIGURE 9. Confusion matrix of proposed model on JAFFE.

scenarios. Confusion matrix of our method on the JAFFE dataset is presented in Figure 9. From the results, it is clearly noticeable that the proposed model attained 100% precision and recall values for angry, neutral and sad emotion classes. Moreover, our method achieved the precision of 85.7% and recall of 100% for both the fear and happy emotions. The drop in performance for fear and happy emotions is attributed to the fact that the JAFFE dataset contains few images where the model subject posed ambiguous facial emotions. These images represent about 5% of the whole dataset.

2) RESULTS ANALYSIS ON CK+ DATASET

To evaluate the performance of our proposed model on the CK+ dataset, we performed three types of experiments on this dataset with varying numbers of emotion classes. Experiment 1 includes seven emotional classes i.e., happy, sad, contempt, fear, anger, disgust, and surprise. Experiment 2 is performed on a total of seven classes in which 6 classes are of basic emotions (happy, fear, sad, surprise, disgust, and anger) and 7th class is neutral. Experiments 1 and 2 differ in terms of the seventh class as experiment 1 contains the contempt emotional class and experiment 2 contains the neutral class. Finally, experiment 3 was performed on the 8 classes including anger (AN), disgust (DI), contempt (CO), fear (FE), happy (HA), sad (SA), surprise (SU), and neutral (NE). For each experiment, we split the dataset into training and test data. Figures 10, 11, and 12 show the progress of our model in terms of training and testing accuracy and loss values for all the three experiments performed on the CK+ dataset.

It is worth noticing from the achieved results presented in form of confusion matrix for all three experiments in Figures 13, 14 and 15 that our model outperforms on all three combinations of classes. For experiments 1 and 2, our model achieved an overall accuracy of 100%. Mean values of precision and recall in these experiments are also 100%. For experiment 3, we obtained the overall accuracy of 95.8%, average precision and average recall of 95.5% and 92.9%, respectively. From the obtained results, it is also obvious that the proposed model also achieved 100% precision and recall value for each class of emotion in both experiments 1 and 2.

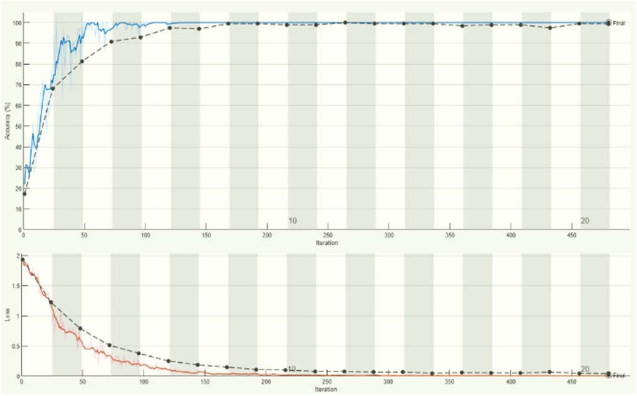


FIGURE 10. Progress of model in basis of accuracy and loss on CK+ dataset for 6 emotion classes + neutral.

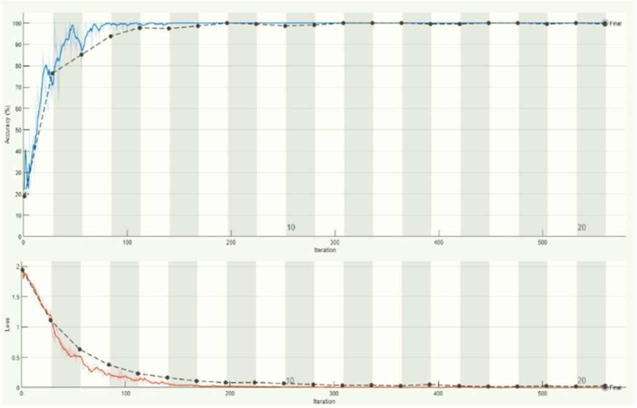


FIGURE 11. Progress of model in terms of accuracy and loss on CK+ dataset for 7 emotion classes.

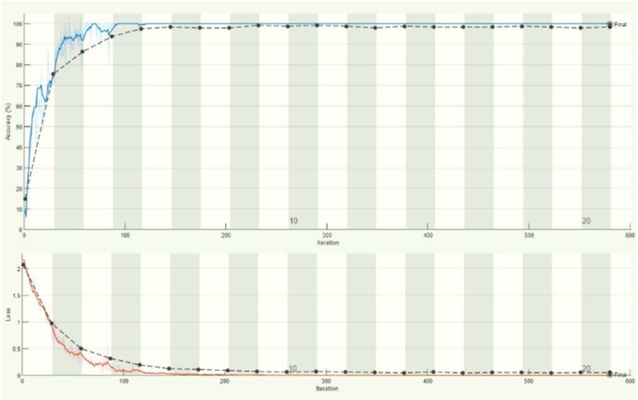
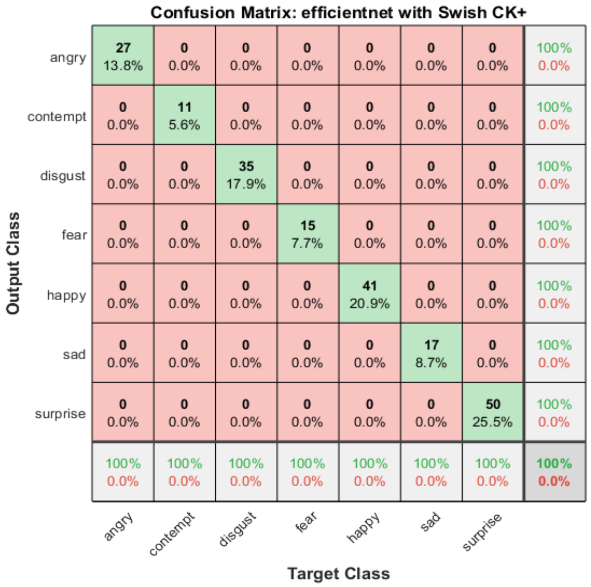


FIGURE 12. Progress of model in terms of accuracy and loss on CK+ dataset for 8 classes.

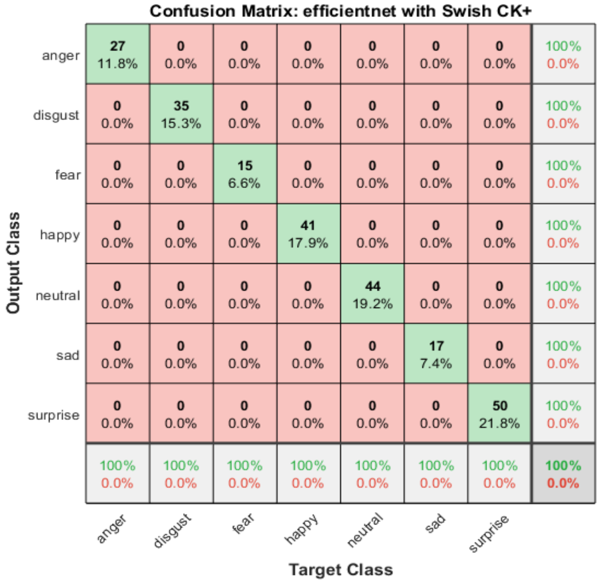


Confusion Matrix: efficientnet with Swish CK+

Output Class / Target Class

	angry	contempt	disgust	fear	happy	sad	surprise	
angry	27 13.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
contempt	0 0.0%	11 5.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
disgust	0 0.0%	0 0.0%	35 17.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
fear	0 0.0%	0 0.0%	0 0.0%	15 7.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
happy	0 0.0%	0 0.0%	0 0.0%	0 0.0%	41 20.9%	0 0.0%	0 0.0%	100% 0.0%
sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	17 8.7%	0 0.0%	100% 0.0%
surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	50 25.5%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

FIGURE 13. Confusion Matrix of proposed model on CK+ for 7 emotion classes.



Confusion Matrix: efficientnet with Swish CK+

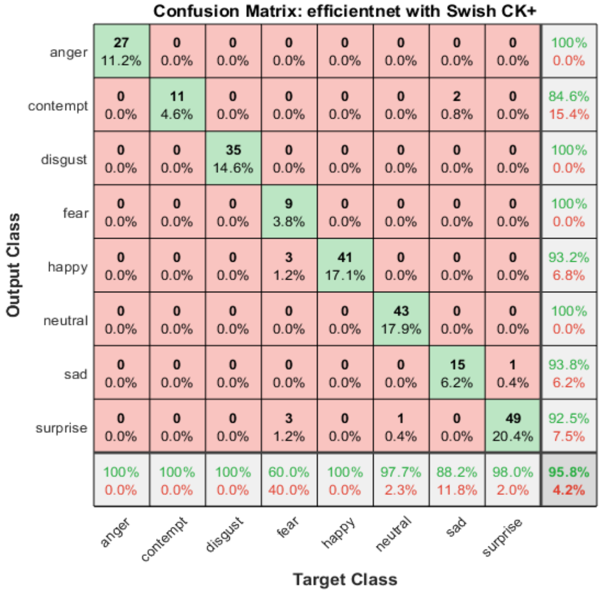
Output Class / Target Class

	anger	disgust	fear	happy	neutral	sad	surprise	
anger	27 11.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
disgust	0 0.0%	35 15.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
fear	0 0.0%	0 0.0%	15 6.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
happy	0 0.0%	0 0.0%	0 0.0%	41 17.9%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	44 19.2%	0 0.0%	0 0.0%	100% 0.0%
sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	17 7.4%	0 0.0%	100% 0.0%
surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	50 21.8%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

FIGURE 14. Confusion Matrix of proposed model on CK+ for 6 emotion classes + neutral.

It can be inferred from results that the proposed model has an ability to distinguish the frontal face emotional images with higher accuracy. But, for experiment 3, recall value for fear class is reduced to 60% as model misclassified few images containing the fear emotion. This might be due to the reason that fear and surprise are closely related emotions. From the results of these three experiments on CK+ dataset, it can also be deduced that Efficient-SwishNet is capable of distinguishing different types of facial emotions from frontal face images with remarkable accuracy of 95-100%.

3) RESULTS ANALYSIS ON KDEF DATASET



Confusion Matrix: efficientnet with Swish CK+

Output Class / Target Class

	anger	contempt	disgust	fear	happy	neutral	sad	surprise	
anger	27 11.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
contempt	0 0.0%	11 4.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.8%	0 0.0%	84.6% 15.4%
disgust	0 0.0%	0 0.0%	35 14.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
fear	0 0.0%	0 0.0%	0 0.0%	9 3.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
happy	0 0.0%	0 0.0%	0 0.0%	3 1.2%	41 17.1%	0 0.0%	0 0.0%	0 0.0%	93.2% 6.8%
neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	43 17.9%	0 0.0%	0 0.0%	100% 0.0%
sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	15 6.2%	1 0.4%	93.8% 6.2%
surprise	0 0.0%	0 0.0%	0 0.0%	3 1.2%	0 0.0%	1 0.4%	0 0.0%	49 20.4%	92.5% 7.5%
	100% 0.0%	100% 0.0%	100% 0.0%	60.0% 40.0%	100% 0.0%	97.7% 2.3%	88.2% 11.8%	98.0% 2.0%	95.8% 4.2%

FIGURE 15. Confusion matrix of proposed model on CK+ for 8 classes.

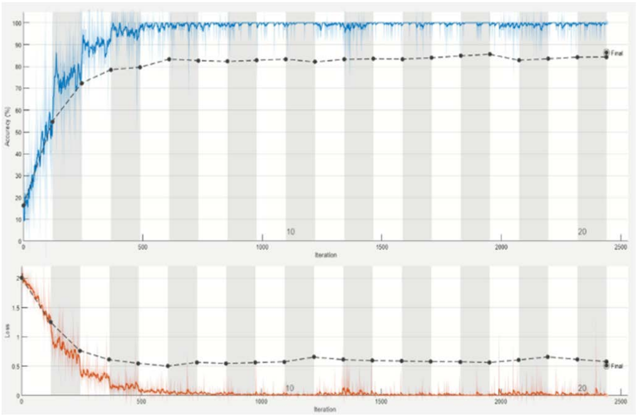


FIGURE 16. Progress of model in terms of accuracy and loss on full KDEF dataset.

To evaluate the robustness of the proposed model for FER on varied angular facial images, we designed two experiments to measure our model's performance on the KDEF dataset as it contains the images of faces captured from different angles. In the first experiment, we selected only the frontal face images as many existing methods use only straight 980 images and in the second experiment, we used the complete dataset with five angles to show that our model can effectively categorize the emotions in images captured from different angles. For both the experiments, the dataset is split into 80:20 ratios (80% for training and 20% for testing). For straight or frontal images, we obtained an overall accuracy of 88.3%, mean precision of 88.3%, and mean recall of 88.2%, whereas achieved an accuracy of 85.5%, mean precision of 85.4%, and mean recall of 85.5% on the overall dataset. This shows that the proposed model is effective in emotions identification on the facial images captured from the frontal as well as tilted angles. Accuracy is dropped to 3% in case of overall dataset with five different viewpoints. This might be due to the fact that in case of full left and full right

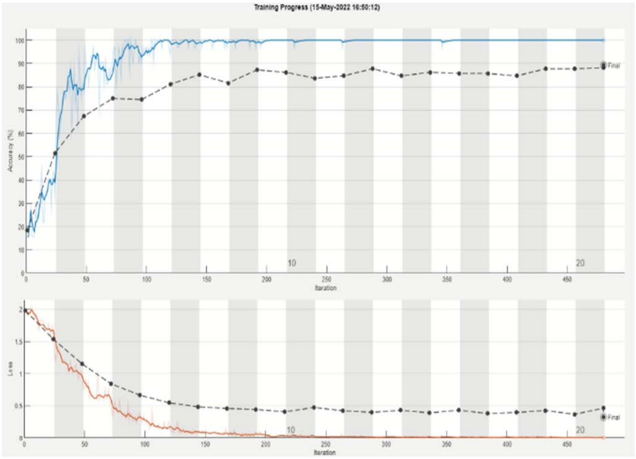
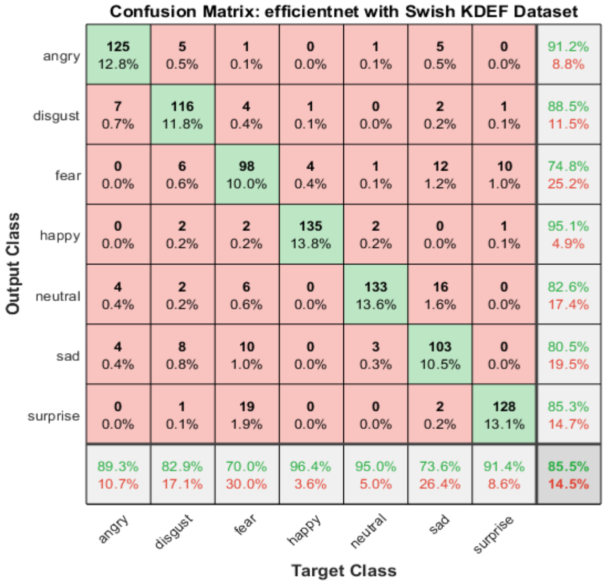


FIGURE 17. Progress of model in terms of accuracy and loss on KDEF straight dataset.


Confusion Matrix: efficientnet with Swish KDEF Dataset

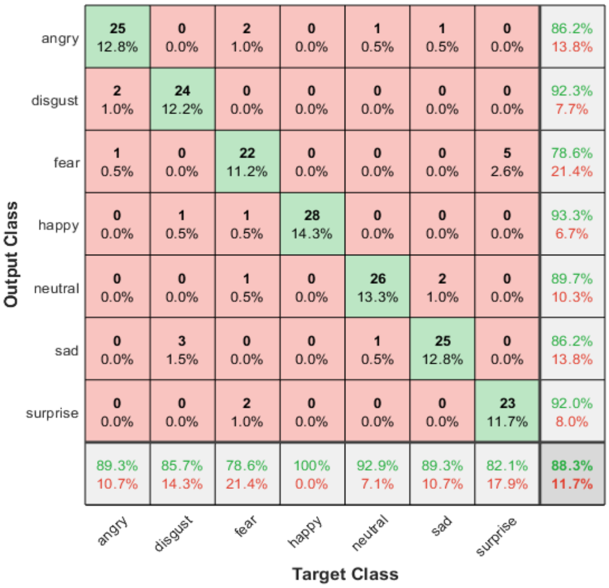
Output Class

Target Class

	angry	disgust	fear	happy	neutral	sad	surprise	
angry	125 12.8%	5 0.5%	1 0.1%	0 0.0%	1 0.1%	5 0.5%	0 0.0%	91.2% 8.8%
disgust	7 0.7%	116 11.8%	4 0.4%	1 0.1%	0 0.0%	2 0.2%	1 0.1%	88.5% 11.5%
fear	0 0.0%	6 0.6%	98 10.0%	4 0.4%	1 0.1%	12 1.2%	10 1.0%	74.8% 25.2%
happy	0 0.0%	2 0.2%	2 0.2%	135 13.8%	2 0.2%	0 0.0%	1 0.1%	95.1% 4.9%
neutral	4 0.4%	2 0.2%	6 0.6%	0 0.0%	133 13.6%	16 1.6%	0 0.0%	82.6% 17.4%
sad	4 0.4%	8 0.8%	10 1.0%	0 0.0%	3 0.3%	103 10.5%	0 0.0%	80.5% 19.5%
surprise	0 0.0%	1 0.1%	19 1.9%	0 0.0%	0 0.0%	2 0.2%	128 13.1%	85.3% 14.7%
	89.3% 10.7%	82.9% 17.1%	70.0% 30.0%	96.4% 3.6%	95.0% 5.0%	73.6% 26.4%	91.4% 8.6%	85.5% 14.5%

FIGURE 18. Confusion matrix of proposed model on overall KDEF dataset.

scenarios one eye and one ear of person's face can be seen, which makes the predictions of different emotions more difficult. Figures 16 and 17 show the progress of our model in the form of accuracy and loss values for experiments performed on the KDEF dataset. We also designed the confusion matrix analysis for both of these experiments and results are shown in Figures 18 and 19.


Output Class

Target Class

	angry	disgust	fear	happy	neutral	sad	surprise	
angry	25 12.8%	0 0.0%	2 1.0%	0 0.0%	1 0.5%	1 0.5%	0 0.0%	86.2% 13.8%
disgust	2 1.0%	24 12.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	92.3% 7.7%
fear	1 0.5%	0 0.0%	22 11.2%	0 0.0%	0 0.0%	0 0.0%	5 2.6%	78.6% 21.4%
happy	0 0.0%	1 0.5%	1 0.5%	28 14.3%	0 0.0%	0 0.0%	0 0.0%	93.3% 6.7%
neutral	0 0.0%	0 0.0%	1 0.5%	0 0.0%	26 13.3%	2 1.0%	0 0.0%	89.7% 10.3%
sad	0 0.0%	3 1.5%	0 0.0%	0 0.0%	1 0.5%	25 12.8%	0 0.0%	86.2% 13.8%
surprise	0 0.0%	0 0.0%	2 1.0%	0 0.0%	0 0.0%	0 0.0%	23 11.7%	92.0% 8.0%
	89.3% 10.7%	85.7% 14.3%	78.6% 21.4%	100% 0.0%	92.9% 7.1%	89.3% 10.7%	82.1% 17.9%	88.3% 11.7%

FIGURE 19. Confusion matrix of proposed model on KDEF straight.

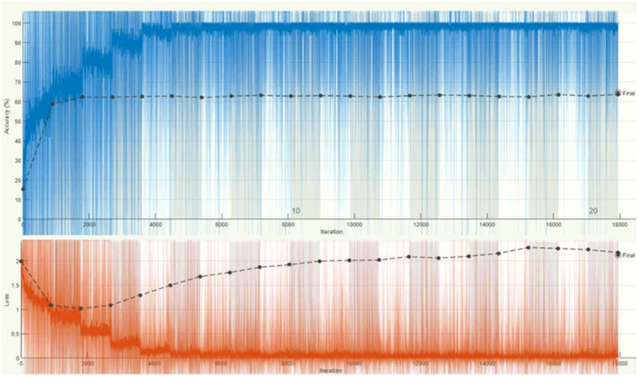
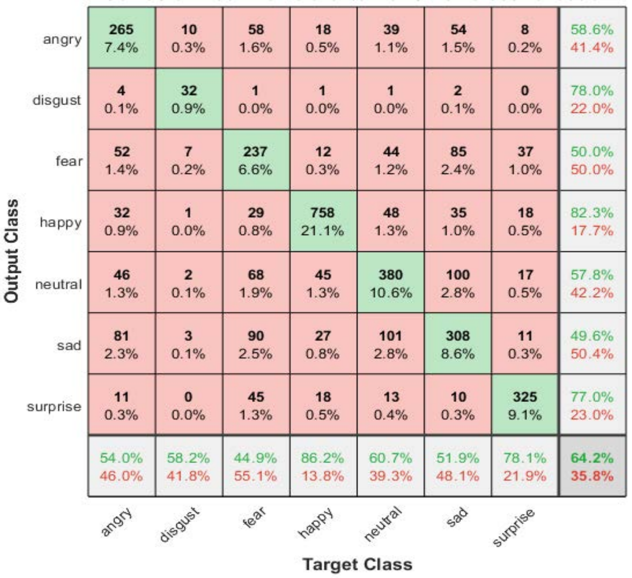


FIGURE 20. Progress of model in terms of accuracy and loss on FER-2013 dataset.

4) RESULTS ANALYSIS ON FER-2013 DATASET

We conducted an experiment to examine the efficacy of the proposed method on a dataset having intra-class variations with class imbalance problem. For this, we selected the real-world FER-2013 dataset that contains such challenging scenarios of intra-class variations and class imbalance. FER-2013 dataset is originally split into training (28,709 images), private test set (3589 images), and public test set or validation set (3589 images). Figure 20 shows the progress of model on FER-2013 dataset. This dataset has intra-class variation with more imbalanced emotional classes, which makes the classification process rather difficult. Moreover, occlusion, non-face images, half-rotated images, varied age and posed images, different illumination conditions, etc., lead to the degradation of accuracy and increased loss.

Figure 21 demonstrates the results of our method on the FER 2013 dataset in the form of a confusion matrix. Our proposed model when tested on a private test set is able to achieve an overall accuracy of 64.2%, mean precision of 64.7%, and mean recall of 62%. The highest precision and recall values are achieved for happy emotion due to a large number of pictures available for this class in the training and testing sets as exhibited in Figure 21. By observing this dataset, it is found that there exist many similarities between the facial morphology of anger and disgust, fear and surprise classes. Moreover, number of images in other classes is far smaller



	angry	disgust	fear	happy	neutral	sad	surprise	
angry	265 7.4%	10 0.3%	58 1.6%	18 0.5%	39 1.1%	54 1.5%	8 0.2%	58.6% 41.4%
disgust	4 0.1%	32 0.9%	1 0.0%	1 0.0%	1 0.0%	2 0.1%	0 0.0%	78.0% 22.0%
fear	52 1.4%	7 0.2%	237 6.6%	12 0.3%	44 1.2%	85 2.4%	37 1.0%	50.0% 50.0%
happy	32 0.9%	1 0.0%	29 0.8%	758 21.1%	48 1.3%	35 1.0%	18 0.5%	82.3% 17.7%
neutral	46 1.3%	2 0.1%	68 1.9%	45 1.3%	380 10.6%	100 2.8%	17 0.5%	57.8% 42.2%
sad	81 2.3%	3 0.1%	90 2.5%	27 0.8%	101 2.8%	308 8.6%	11 0.3%	49.6% 50.4%
surprise	11 0.3%	0 0.0%	45 1.3%	18 0.5%	13 0.4%	10 0.3%	325 9.1%	77.0% 23.0%
	54.0% 46.0%	58.2% 41.8%	44.9% 55.1%	86.2% 13.8%	60.7% 39.3%	51.9% 48.1%	78.1% 21.9%	64.2% 35.8%

FIGURE 21. Confusion matrix of proposed model on FER-2013.

than the number of happy images, which leads to insufficient learning of features for these classes, so the recognition rate is low for angry, fear and sad classes.

5) RESULTS ANALYSIS ON FERG DATASET

To evaluate the effectiveness of the proposed model for FER on cartoon characters, we designed an experiment to measure our model's performance on the FERG dataset as it contains the images of cartoonish characters. We split the dataset into training and test sets. The training dataset includes 44,613 images (80% approx.) while the testing set contains 11,154 images (20% approx.). Our Efficient-SwishNet achieved an overall accuracy, mean precision, and mean recall of 100% for the FERG dataset which shows the effectiveness of our model for emotion recognition in the case of cartoon characters. Figure 22 shows the progress of our model on the FERG stylish cartoon characters' dataset. The confusion matrix of our method on the FERG dataset is displayed in Figure 23. From the results, it can be clearly observed that our proposed model obtained the optimal results on this dataset where none of the images is misclassified by the proposed model.

Table 2 shows the summarized performance of Efficient-SwishNet in terms of overall accuracy, mean precision, mean recall, mean specificity, and macro F-1 score on five diverse datasets involved in the experimentation. KDEF dataset contains the images from five different angles, CK+ dataset contains the images of people belonging to different geographical regions. JAFFE is a small and biased dataset for gender as only female Japanese models are involved in the preparation of this dataset. Overall, all these datasets have varied illumination conditions, people belong to different regions and continents of the world with different races. Significant results on these datasets proved that our model is robust to changes in face angles, illumination conditions, gender, different background settings, and morphological appearance of people belonging to diverse geographical regions on the posed as well as spontaneous images and capable of distinguishing different types of emotions precisely. Moreover, the proposed model not only identify the emotions of human with higher accuracy but is also capable of distinguishing the emotions of stylized cartoon characters.

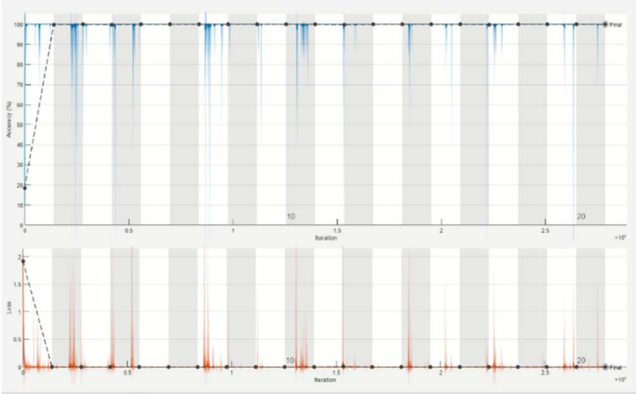
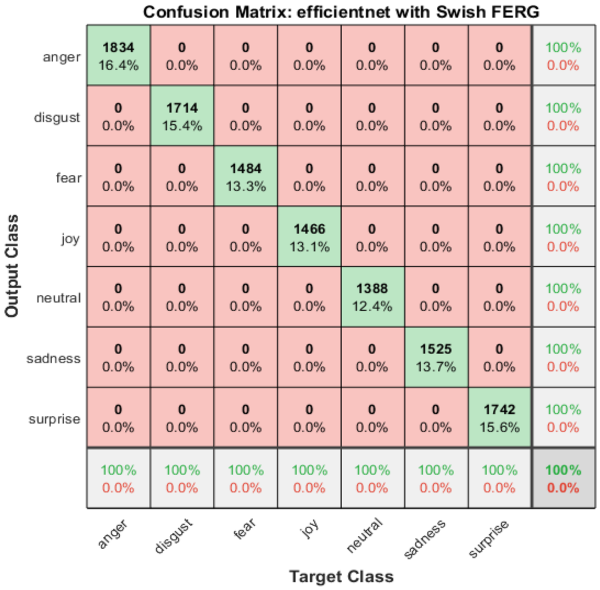


FIGURE 22. Progress of model in terms of accuracy and loss on FERG dataset.



	anger	disgust	fear	joy	neutral	sadness	surprise	
anger	1834 16.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
disgust	0 0.0%	1714 15.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
fear	0 0.0%	0 0.0%	1484 13.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
joy	0 0.0%	0 0.0%	0 0.0%	1466 13.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1388 12.4%	0 0.0%	0 0.0%	100% 0.0%
sadness	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1525 13.7%	0 0.0%	100% 0.0%
surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1742 15.6%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

FIGURE 23. Confusion matrix of proposed model on FERG.

D. COMPARATIVE STUDY

To show the effectiveness of our model for the recognition of facial emotion on multiple diverse datasets, we performed a multi-stage experiment to compare the performance of our method against the previous state-of-the-art (SOTA) FER methods.

In the first stage, we compared the performance of our method on these contemporary methods [1], [13], [14], [16], [17], and [22] on the JAFFE dataset and results in terms

TABLE 2. Performance of the proposed model on different datasets.

Dataset	Overall Accuracy	Mean Precision	Mean Recall/Sensitivity	Mean Specificity	Macro F-1 score
JAFFE	95.02%	95.9%	95.2%	0.98	0.951
CK+ (Experiment 1)	100%	100%	100%	1.0	1.0
CK+ (Experiment 2)	100%	100%	100%	1.0	1.0
CK+ (Experiment 3)	95.8%	95.5%	92.9%	0.99	0.935
KDEF (Straight images)	88.3%	88.3%	88.2%	0.97	0.884
KDEF (full dataset)	85.5%	85.5%	85.4%	0.97	0.852
FER-2013 dataset	64.2%	64.7%	64.2%	0.93	0.631
FERG dataset	100%	100%	100%	1.0	1.0

TABLE 3. Comparative analysis on JAFFE dataset.

Model	Accuracy
Minaee et al. [1]	92.8%
Sun et al. [22]	61.68%
Khaliluzzamanet al. [13]	93.02%
Dubey et al. [14]	93.75%
Kola et al. [17]	88.3%
LBP+ORB [16]	92.4%
Jain et al. [40]	95.23%
Proposed Model	**95.02%**

TABLE 4. Comparative analysis on CK+.

Model	Classes	Accuracy
Minaee et al. [1]	7 classes (AN,HA,SA,CO,FE,SU, DI)	98%
LBP+ORB [16]	7 classes (AN,HA,SA,CO,FE,SU, DI)	99.2%
ResNet50 [38]	7 classes (AN,HA,SA,CO,FE,SU, DI)	97.7%
Inception-V3 [38]	7 classes (AN,HA,SA,CO,FE,SU, DI)	94.2%
Mobile Net [38]	7 classes (AN,HA,SA,CO,FE,SU, DI)	98.5%
VGG 19 [38]	7 classes (AN,HA,SA,CO,FE,SU, DI)	96.0%.
Proposed Model	**7 classes (AN,HA,SA,CO,FE,SU, DI)**	**100%**
Khaliluzzaman et al. [13]	7 classes (AN,HA,SA,NE,FE,SU, DI)	99.49%
Zhong et al. [15]	7 classes (AN,HA,SA,NE,FE, SU, DI)	95.253 %
Jain et al. [40]	7 classes (AN, HA, SA, NE, FE, SU, DI)	93.24%
Proposed Model	**7 classes (AN,HA,SA,NE,FE,SU, DI)**	**100%**
Dubey et al. [14]	8 classes (AN,HA,SA,CO,FE,SU, DI,CO)	94.84%
Proposed Model	**8 classes (AN,HA,SA,CO,FE,SU, DI,CO)**	**95.8%**

TABLE 5. Comparative analysis on KDEF dataset.

Model	Accuracy
Standard CNN (224x224 size images) [20]	73.46%
STL + RTNN [21]	84.08%
RTNN + Laplacian RTNN [21]	84.39%
Sun et al. [22]	77.98%
Proposed Model (full dataset)	**85.5%**
HOG with SVM [23]	80.2 ± 4.1
HOG with Adaboost [23]	75.2 ± 4.0
RES + Avgpool + SVM [33]	76.8%
RES + 49Relu + SVM [33]	83.9%
RES + Add16 + SVM [33]	82.3%
Proposed Model (frontal Images)	**88.3%**

of accuracy are provided in Table 3. The proposed method provided superior detection accuracy over the comparative methods. More precisely, our method obtained an accuracy of 95.02%, while the second-best performing method [14] obtained an accuracy of 93.75%. This might be due to the fact that [14] used the ReLU activation function which has a comparatively low recognition rate than the Swish activation function. [22] achieved the lowest accuracy of 61.68%. Our method achieved an average accuracy gain of 8% over the comparative methods. These results show the effectiveness of the proposed method for FER on the JAFFE dataset over the contemporary methods.

In the second stage, we designed three different experiments based on a different number of emotion classes as discussed in *Section V-C2* to compare our method with the existing approaches on the CK+ dataset. It is worth noticing that the proposed model correctly classifies all the testing images in experiments 1 and 2 as compared to [1], [13], [15], and [16]. For experiment 3 with eight emotional classes, our Efficient-SwishNet model obtained an accuracy greater than the comparative method [14]. The reported accuracy of the proposed model is higher than all of the contemporary works for all three experiments performed with a different number of classes as shown in Table 4.

In the third stage, we designed two different experiments with frontal face images and on the entire dataset as discussed in *Section V-C3* to compare our model with prior works on KDEF dataset. Table 5 signifies the effectiveness of our model over these comparative works [20-23], [33]. It can be seen that the proposed method achieves higher discriminative ability for different emotions than these methods [20-22] for complete dataset having images captured from five different angles. Moreover, for straight images, recognition rate is increased to 10-15% than the comparative approaches [23] and [33].

In the fourth stage, we compared our proposed model on FER-2013 dataset with these previous state-of-the-art methods [1], [4], [8] and [29]. Our Efficient-SwishNet model not only classifies the frontal or angled images (on CK+, JAFFE, and KDEF datasets) with greater accuracy but also performs well as compared to many existing approaches on FER-2013 dataset as presented in Table 6. This attribute of the proposed method enables it to be implemented in real-world applications for facial emotion detection.

In the last stage of this experiment, we compared the performance of our method on the FERG character database with these approaches [1], [17], and [19] as shown in Table 7.

TABLE 6. Comparative analysis on FER-2013 dataset.

Model		Accuracy
Talegaonkar et al. [8]		60.12%
Ramdhani et al. [4]	With batch size 8	58.20%
	With batch size 128	62.33%
Xception without FIT machine et al. [29]		60.99%
Xception with FIT machine et al. [29]		63.99%
Minaee et al. [1]		70.02%
Proposed Model		**63.4%**

TABLE 7. Comparative analysis on FERG dataset.

Model	Accuracy
Minaee et al. [1]	99.3%
Kola et al. [17]	96.7%
DeepExpr [19]	89.02%
Proposed Model	**100%**

Our model not only performed well for the identification of human images but also achieved a 100% recognition rate for the FERG stylized character dataset which none of the existing FER methods could achieve till now. Our method not only shows good detection results than existing SOTA works on small and posed datasets but also shows 100% accuracy on a largescale and diverse FERG dataset. This shows the effectiveness of our Efficient-SwishNet model for the reliable detection of emotions on cartoon faces.

In general, our proposed method provides remarkable detection results on all five datasets against the prior SOTA methods which show its ability to detect seven or eight different types of facial emotions. Our method also covered the limitations of prior approaches of different illumination conditions, viewpoints, detection of few numbers of emotions, gender, people of different areas, and race. We used five diverse datasets for the evaluation of our proposed method, which are completely different from each other and contains the facial images with diverse scenarios and images of humans as well as of cartoon faces. From this remarkable detection performance of our model than the existing methods on these datasets, it is obvious that our model is not limited to the detection of emotions for only straight facial images but is also capable of detecting the emotions in the case of angled face images and detect emotions of people belonging to different continents, race, and gender. Furthermore, our method can reliably be used to detect facial emotions while achieving better efficiency.

E. ABLATION STUDY

In this section, we conducted three different experiments. One experiment is designed to evaluate the efficacy and efficiency of our proposed model over existing DL models in terms of accuracy, number of parameters, and execution time. While the other two experiments are designed to check the effect of different activation functions on the accuracy and efficiency of our proposed model. Evaluation is done on the CK+ dataset with 7 emotional classes and experimental protocols for these experiments are also kept the same as done in experiment 1 described in *Section V-C2*. All experiments of ablation study are carried out on a machine having specifications of 32GB Ram, AMD Ryzen 9 5900x 12-core 3.70GHz processor, 4.5TB hard disk, and Windows10 Pro. The details of these experiments are mentioned in the subsequent sections.

TABLE 8. Time complexity analysis of Proposed and comparative models.

Existing CNN models	Acc.	Execution time	Model Size	#Param
DenseNet-201	100%	1360s ≈ 22.6min	77MB	20.0M
InceptionResnetv2	99.9%	1120s ≈ 18.6min	209MB	55.9M
Inception-v3	97.5%	393s ≈ 6.5min	89MB	23.9M
NasNet-Mobile	99.5%	1625s ≈ 27min	20MB	5.30M
EfficientNet with Swish (proposed)	100%	505s ≈ 8.40min	19.9MB	5.31M

1) EFFCIENCY COMPARISON OF EFFICIENT-SWISHNET WITH EXISTING CNN MODELS

We conducted an experiment to prove that our model is computationally more efficient than other deep learning models. Table 8 presents the comparison results of Efficient-SwishNet with existing models in terms of number of parameters, accuracy, size of model, and time complexity. Results show that Efficient-SwishNet has achieved the maximum accuracy of 100% with the fewest parameters and minimum space requirement and minimum execution time. The DenseNet-201 also achieved the accuracy of 100%, however, at the expense of increased execution time and number of parameters. More specifically, our model uses 14.69M and 50.59M lesser number of parameters than the DenseNet-201 and InceptionResnet-v2 models. Execution time for Inception-v3 is 1.9 minutes lesser but its accuracy is reduced to 2.5% and number of parameters is 18.59M more than the Efficient-SwishNet. Similarly, NasNet-Mobile is computationally the most expensive model with an execution time of 27 minutes and accuracy is 0.5% lesser than our proposed model. Furthermore, our proposed model is also space-efficient than the comparative CNN models as shown in Table 8. From this experiment, it is proved that our proposed model is not only computationally efficient but also space-efficient than the models mentioned in Table 8.

2) IMPACT OF THE ACTIVATION FUNCTION ON MODEL'S ACCURACY

To show the effect of different activation functions on the accuracy of our custom EfficientNet-b0 architecture, we conducted few experiments with the original EfficientNet-b0 model, EfficientNet-b0 with ReLU, Leaky-ReLU, ELU, Sigmoid, tanh, and Swish activation functions.

The results of this comparative analysis of different activation functions are stated in Table 9. From these results, we can clearly see that our proposed model outperforms all the variants and original EfficientNet-b0 by achieving 100% optimal recognition rate. EfficientNet-b0 with the ReLU activation

TABLE 9. Results of different variants of proposed model.

Variants of EfficientNet-b0	Dataset	Accuracy
EfficientNet-b0	CK+	79.01%
EfficientNet-b0 with ReLU	CK+	84.02%
EfficientNet-b0 with Leaky-ReLU	CK+	71.6%
EfficientNet-b0 with ELU	CK+	71.6%
EfficientNet-b0 with Sigmoid	CK+	67.48%
EfficientNet-b0 with tanh	CK+	44.3%
EfficientNet with swish (proposed)	CK+	100%

TABLE 10. Time complexity analysis of different variants of proposed model.

Variants of EfficientNet-b0	Dataset	Execution time
EfficientNet-b0	CK+	525s ≈ 08.75 min
EfficientNet-b0 with ReLU	CK+	804s ≈ 13.40 min
EfficientNet-b0 with Leaky-ReLU	CK+	526s ≈ 08.76 min
EfficientNet-b0 with ELU	CK+	545s ≈ 09.00 min
EfficientNet-b0 with Sigmoid	CK+	522s ≈ 08.70 min
EfficientNet-b0 with tanh	CK+	519s ≈ 08.60 min
EfficientNet with swish (proposed)	CK+	498s ≈ 08.30min

performed second best over all the other variants, whereas we found that the EfficientNet-b0 with tanh activation function has achieved the lowest performance. Number of parameters for all the variants is kept same i.e., 5.31M. Table 9 illustrated that our Efficient-SwishNet model can reliably be used to recognize various facial emotions on a diverse and challenging dataset.

3) IMPACT OF ACTIVATION FUNCTION ON MODEL EFFICIENCY

We performed an experiment to analyze the execution time of different activation functions to show that the proposed model is more efficient and low-cost than its variants. For this purpose, we employed six activation functions i.e., ELU, leaky ReLU, ReLU, tanh, sigmoid, and Swish in EfficientNet-b0 architecture and also considered the original EfficientNet-b0 model to analyze the computational cost. Table 10 shows the execution time of different variants of EfficientNet-b0 used in our research work.

From Table 10, it is inferred that ReLU is computationally the most expensive activation function as it requires the maximum execution time. Swish has the least execution time and achieved the highest accuracy as compared to other functions. So, we can say that Swish is low-cost and more accurate activation function. Therefore, the introduction of Swish in our customized EfficientNet-b0 architecture makes the proposed model computationally more efficient and lightweight.

F. CROSS CORPORA EVALUATION

To check the generalizability of the proposed model, we conducted a cross corpora evaluation as almost none of the previous works on facial emotion recognition gave attention to the aspect of model generalization for seven or eight classes of emotions. All the datasets are first split into an 80:20 ratio (80% for training and 20% for testing). The model is trained on one dataset and tested on other datasets. Results of cross-corpus experiments are stated in Table 11 and results above 30% are shown in bold. Despite the excellent performance of our model on individual datasets, it could not perform very well on cross-corpus experiments. This is due to the reason that these datasets are collected under different environments with different illumination conditions, varying distances from the camera, and background settings.

TABLE 11. Cross corpora evaluation results.

Training	Testing	Accuracy
JAFFE	CK+	10.5%
	KDEF full	16.6%
	KDEF straight	16.8%
	FERG	13.4%
	FER-2013	13.1%
CK+	JAFFE	11.9%
	KDEF full	14.6%
	KDEF straight	16.3%
	FERG	14.8%
	FER-2013	23.8%
FERG	CK+	11.4%
	KDEF full	18.0%
	KDEF straight	**30.1%**
	JAFFE	**38.1%**
	FER-2013	13.1%
FER-2013	**CK+**	**58.1%**
	KDEF full	17.6%
	KDEF straight	23.5%
	JAFFE	16.7%
	FERG	**30.9%**
KDEF	CK+	19.7%
	JAFFE	23.8%
	FER-2013	16.6%
	FERG	14.0%

Moreover, individuals involved in the making of these datasets are not of the same age group, or gender and belong to different geographical regions (JAFFE: Asia, KDEF: Europe and CK+: mostly subjects are African American) of the world. There also exists a dissimilarity in emotional expression between different cultures, western people tend to experience high arousal emotions as compared to collectivist or Eastern culture [30]. Furthermore, Japanese participants, in contrast to Americans and Europeans, reported significantly fewer physiological emotions [30]. Subjects involved in the preparation of these datasets have a huge diversity in morphological characteristics. We also included a stylized characters FERG dataset in cross-validation experiment which is one of the largescale available datasets for FER, but this dataset varies from other datasets in the aspect that it contains only the cartoon facial images and other datasets contain human images which makes the cross corpora evaluation more challenging between such different types of datasets. Despite all these reasons, it can be observed that our model trained on the FER-2013 dataset and tested on the CK+ dataset attained an accuracy of 58%. These results are encouraging considering such diversity in these datasets. Similarly, the model trained on the FERG dataset and tested on the JAFFE dataset achieved an accuracy of 38.1%.

VI. CONCLUSION

In this paper, we have presented a novel Efficient-SwishNet model, which is cost-effective, robust, and space-efficient for facial emotion recognition. The proposed model is evaluated on five different datasets and outperforms the state-of-the-art methods on all the datasets. The proposed model not only accurately classified the emotions from frontal face images but also performed well than existing methods for face images with five different orientations. Furthermore, the recognition rate on FERG and CK+ datasets is 100%, which shows that the proposed model is able to distinguish all the emotion classes accurately and can be implemented in real-world applications. We also performed cross corpora evaluation of the proposed model to show the generalizability of our method. Moreover, different variants of the proposed model are also evaluated which shows that the proposed method is more efficient and robust in the identification of facial emotions. It is important to mention that our model is unable to perform well in case of occlusion or when the face is covered with accessories like glasses, masks, hair, etc. Furthermore, we did not obtain much better results in the case of cross corpora evaluation. In the future study, we aim to develop a FER model capable of achieving improved performance for cross corpora evaluation and covered faces with occlusions. Furthermore, we also plan to create a custom FER dataset to test the performance of our future method in real-time.

REFERENCES

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

[2] A. N. Dixit and T. Kasbe, "A survey on facial expression recognition using machine learning techniques," in *Proc. 2nd Int. Conf. Data, Eng. Appl. (IDEA)*, Feb. 2020, pp. 1–6.

[3] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *ELCVIA Electron. Lett. Comput. Vis. Image Anal.*, vol. 14, no. 2, pp. 24–44, Dec. 2015.

[4] B. Ramdhani, E. C. Djamal, and R. Ilyas, "Convolutional neural networks models for facial expression recognition," in *Proc. Int. Symp. Adv. Intell. Informat. (SAIN)*, Aug. 2018, pp. 96–101.

[5] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using CNN architecture," in *Proc. Med. Technol. Congr. (TIPTEKNO)*, Oct. 2019, pp. 1–4.

[6] S. Pande and S. Shinde, "A survey on: Emotion recognition with respect to database and various recognition techniques," *Int. J. Comput. Appl.*, vol. 58, no. 3, pp. 9–12, Nov. 2012.

[7] H. Yang, G. Zhao, L. Zhang, N. Zhu, Y. He, and C. Zhao, "Real-time emotion recognition framework based on convolution neural network," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing* (Smart Innovation, Systems and Technologies). Singapore: Springer, 2020, pp. 313–321.

[8] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real time facial expression recognition using deep learning," in *Proc. Int. Conf. Commun. Inf. Process. (ICCIP)*, 2019, pp. 1–8.

[9] I. Cugu, E. Sener, and E. Akbas, "MicroExpNet: An extremely small and fast model for expression recognition from face images," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2019, pp. 1–6.

[10] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," 2017, *arXiv:1710.07557*.

[11] H. I. Yar, T. A. Jan, A. L. Hussain, and S. U. Din, "Real-time facial emotion recognition and gender classification for human robot interaction using CNN," in *Proc. 5th Int. Conf. Next Gener. Comput.*, 2019, pp. 20–21.

[12] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *Social Netw. Appl. Sci.*, vol. 2, no. 3, pp. 1–8, Mar. 2020.

[13] M. Khaliluzzaman, S. Pervin, M. R. Islam, and M. M. Hassan, "Automatic facial expression recognition using shallow convolutional neural network," in *Proc. IEEE Int. Conf. Robot., Autom., Artif.-Intell. Internet-Things (RAAICON)*, Nov. 2019, pp. 98–103, doi: 10.1109/RAAICON48939.2019.42.

[14] A. K. Dubey and V. Jain, "Automatic facial recognition using VGG16 based transfer learning model," *J. Inf. Optim. Sci.*, vol. 41, no. 7, pp. 1589–1596, Oct. 2020.

[15] Y. Zhong, S. Qiu, X. Luo, Z. Meng, and J. Liu, "Facial expression recognition based on optimized ResNet," in *Proc. 2nd World Symp. Artif. Intell. (WSAI)*, Jun. 2020, pp. 84–91.

[16] B. Niu, Z. Gao, and B. Guo, "Facial expression recognition with LBP and ORB features," *Comput. Intell. Neurosci.*, vol. 2021, Jan. 2021, Art. no. 8828245.

[17] D. G. Kola and S. K. Samayamantula, "A novel approach for facial expression recognition using local binary pattern with adaptive window," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2243–2262, Jan. 2021.

[18] K. Ounachad, "Geometric feature based facial emotion recognition," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3417–3425, Jun. 2020.

[19] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, Nov. 2016, pp. 136–153.

[20] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, Apr. 2021, doi: 10.3390/electronics10091036.

[21] R. K. Pandey, S. Karmakar, A. G Ramakrishnan, and N. Saha, "Improving facial emotion recognition systems using gradient and Laplacian images," 2019, *arXiv:1902.05411*.

[22] Z. Sun, Z.-P. Hu, M. Wang, and S.-H. Zhao, "Individual-free representation-based classification for facial expression recognition," *Signal, Image Video Process.*, vol. 11, no. 4, pp. 597–604, May 2017.

[23] C. F. Liew and T. Yairi, "Facial expression recognition and analysis: A comparison study of feature descriptors," *IPSJ Trans. Comput. Vis. Appl.*, vol. 7, pp. 104–120, Apr. 2015.

[24] M. Murugappan and A. Mutawa, "Facial geometric feature extraction based emotional expression classification using machine learning algorithms," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0247131.

[25] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces (KDEF)," CD ROM Dept. Clinical Neurosci., Psychol. Sect., Karolinska Institutet, Solna, Sweden, Tech. Rep., 1998, vol. 91, no. 630.

[26] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, and Y. Zhou, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, Nov. 2013, pp. 117–124.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[28] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[29] J. H. Kim, A. Poulose, and D. S. Han, "The extensive usage of the facial image threshing machine for facial emotion recognition performance," *Sensors*, vol. 21, no. 6, p. 2026, Mar. 2021.

[30] N. Lim, "Cultural differences in emotion: Differences in emotional arousal level between the East and the West," *Integrative Med. Res.*, vol. 5, no. 2, pp. 105–109, Jun. 2016.

[31] K. R. Scherer, D. Matsumoto, H. G. Wallbott, and T. Kudoh, "Emotional experience in cultural context: A comparison between Europe, Japan and the United States," in *Facets of Emotion: Recent Research*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1998, pp. 98–115.

[32] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets (IVC special issue)," 2020, *arXiv:2009.05938*.

[33] Z. Fei, E. Yang, D. Li, S. Butler, W. Ijomah, and H. Zhou, "Combining deep neural network with traditional classifier to recognize facial expressions," in *Proc. 25th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2019, pp. 1–6.

[34] M. N. Islam and C. K. Loo, "Geometric feature-based facial emotion recognition using two-stage fuzzy reasoning model," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, Nov. 2014, pp. 344–351.

[35] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm, "Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1013–1022.

[36] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.

[37] P. Jiang, B. Wan, Q. Wang, and J. Wu, "Fast and efficient facial expression recognition using a Gabor convolutional network," *IEEE Signal Process. Lett.*, vol. 27, pp. 1954–1958, 2020.

[38] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 33, pp. 1–8, Apr. 2021.

[39] E. Pranav, S. Kamal, C. S. Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 317–320.

[40] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

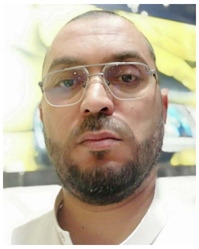
[42] *Maya: Create Expansive Worlds, Complex Characters, and Dazzling Effects*. Accessed: May 20, 2022. [Online]. Available: https://www.autodesk.com/products/maya/overview?term=1-YEAR&tab=subscription&plc=MAYA

TARIM DAR received the bachelor's degree in software engineering from the UET Taxila, Pakistan, in 2019, where she is currently pursuing the M.S. degree with the Software Engineering Department. Her research interests include deep learning, audio image, and video processing, and machine learning.

ALI JAVED (Member, IEEE) received the B.Sc. degree (Hons.) in software engineering and the M.S. and Ph.D. degrees in computer engineering from the UET Taxila, Pakistan, in 2007, 2010, and 2016, respectively.

He is currently working as an Associate Professor with the Software Engineering Department, UET Taxila. Previously, he worked as an Assistant Professor with the Software Engineering Department, UET Taxila. He worked as the HOD with the Software Engineering Department, UET Taxila, in 2014. He worked as a Postdoctoral Scholar with the SMILES Laboratory, Oakland University, USA, in 2019, and as a Visiting Ph.D. Scholar with the ISSF Laboratory, University of Michigan, USA, in 2015. His research interests include digital image processing, computer vision, video content analysis, machine learning, multimedia signal processing, and multimedia forensics.

Dr. Javed has been a member of the Pakistan Engineering Council, since 2007. He was a recipient of various research grants from HEC Pakistan, National ICT Research and Development Fund, NESCOM, and UET Taxila. He received the Chancellor's Gold Medal in M.S. degree in computer engineering. He got selected as an Ambassador of Asian Council of Science Editors from Pakistan, in 2016.



SAMI BOUROUIS received the Engineer, M.Sc., and Ph.D. degrees in computer science from the University of Tunis, Tunisia, in 2003, 2005, and 2011, respectively. He is currently an Associate Professor at the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include data mining, image processing, statistical machine learning, cybersecurity, and pattern recognition applied to several real-life applications.

HANY S. HUSSEIN (Senior Member, IEEE) received the B.Sc. degree in electrical engineering and the M.Sc. degree in communication and electronics from South Valley University, Egypt, in 2004 and 2009, respectively, and the Ph.D. degree in communication and electronics engineering from the Egypt–Japan University of Science and Technology (E–JUST), in 2013. In 2012, he worked as a Special Researcher Student at Kyushu University, Japan. He has been an Associate Professor with the Faculty of Engineering, Aswan University, since 2019. He is currently working as an Assistant Professor with the College of Engineering, King Khalid University, Saudi Arabia. His research interests include digital signal processing for communications, multimedia, image, and video coding, low-power wireless communications, one-bit ADC multiple-input multiple-output, underwater communication, index, and spatial modulation, Li-Fi technology, and visible light communication. He is a technical committee member of many international conferences and a reviewer of many international conferences, journals, and transactions. Moreover, he was the General Co-Chair of the IEEE ITCE, in 2018.

HAMMAM ALSHAZLY received the B.Sc. degree in computer science from South Valley University, Egypt, in 2006, the M.Sc. degree in computer science from the University of Mumbai, India, through a scholarship from the Indian Council for Cultural Relations (ICCR), in 2014, and the Ph.D. degree in computer science from South Valley University, in 2018. He is currently working as an Assistant Professor with the Department of Computer Science, Faculty of Computers and Information, South Valley University. From February 2019 to January 2021, he was a Postdoctoral Researcher with the Institute for Neuro- and Bioinformatics, University of Lübeck, Germany. He has published papers in conferences and peer-reviewed journals and works as a reviewer for several journals. His research interests include deep learning, biometrics, computer vision, machine learning, and artificial intelligence.