Hybrid Deep Neural Network for Facial Expressions Recognition

Wijdan R. Abdulhussein[1], Nidhal K. El Abbadi[2], Abdul M. Gaber[3]
[1]Department of Computer Science, University of Technology, Baghdad, Iraq.
[2]Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq.
[2]Turath University College, Baghdad, Iraq.

Article Info

Article history:

Received Oct 25, 2021
Revised Nov 26, 2021
Accepted Dec 14, 2021

Keywords:

Facial Expression Recognition
Transfer learning
Convolutional neural network
Facial emotion
Fer2013 Dataset

ABSTRACT

Facial expressions are critical indicators of human emotions where recognizing facial expressions has captured the attention of many academics, and recognition of expressions in natural situations remains a challenge due to differences in head position, occlusion, and illumination. Several studies have focused on recognizing emotions from frontal images only, while in this paper wild images from the FER2013 dataset have been used to make a more generalizing model with the existence of its challenges, it is among the most difficult datasets that only got 65.5 % accuracy human-level. This paper proposed a model for recognizing facial expressions using pre-trained deep convolutional neural networks and the technique of transfer learning. this hybrid model used a combination of two pre-trained deep convolutional neural networks, training the model in multiple cases for more efficiency to categorize the facial expressions into seven classes. The results show that the best accuracy of the suggested models is 74.39% for the hybrid model, and 73.33% for Fine-tuned the single EfficientNetB0 model, while the highest accuracy for previous methods was 73.28%. Thus, the hybrid and single models outperform other state of art classification methods without using any additional, the hybrid and single models ranked in the first and second position among these methods. Also, The hybrid model has even outperformed the second-highest in accuracy method which used extra data. The incorrectly labeled images in the dataset unfairly reduce accuracy but our best model recognized their actual classes correctly.

Corresponding Author:

Wijdan R. Abdulhussein,
Department of Computer Science, University of Technology,
Technology University,
Baghdad, Iraq.
Email: wijdan_rashid@utq.edu.iq

1. INTRODUCTION

One of the strongest, natural, and global signals that reflect human emotions and state is facial expression. In the age of artificial intelligence, facial expression recognition (FER) is essential. Machines can give customized services based on human reaction data. Many systems, such as personalized suggestions, virtual reality, customer satisfaction, etc., rely on the ability to identify facial expressions quickly and accurately.In recent years, due to the rapid growth in artificial intelligence, automatic facial expression recognition (FER) has an increasing interest among researchers in the field of computer vision, psychology, and pattern recognition [1], also, facial expression's classification has become a fundamental portion of computer systems and the quick interaction between humans and computers. It is used in many applications, including customer service, forensic crime detection, observation of victims in court, mentoring of students in academics [2], robotics, digital entertainment, advanced driver assistance, and monitoring systems [3], virtual reality, augmented reality, and education. Although many researchers work on developing the robust FER, they still face many problems and challenges that reduce the recognition accuracy rate, such as illumination, noise, and occlusion that affect the feature extraction. In addition, the large data dimension can deteriorate the accuracy rate and recognition performance. Deep learning solved some of the problems but still needs a large amount of data to overcome the overfitting. In addition, many other parameters have a significant effect on the recognition performance, some of them are a variation of personal attributes that can cause high inter-subject variations, in addition to age, gender, ethnic backgrounds, and pose variation [4].

In this paper, we will be focusing on only improving the recognition of facial expression using the FER2013 dataset without adding any further data, while most similar works done involves adding extra training data from other sources and datasets to improve the accuracy rather than work on the recognition model itself, and thus compare the current work with works of other researches.

The rest of this paper is organized as follows: Section two focuses on the facial expression analysis, followed by related works in section three. While section four introduced a general background, about three of the main methods used in this work (CNN, Deep CNN, and transfer learning). Then we covered the proposed model in section five. Results and performance comparing have been discussed in sections six and seven. Finally, the discussion and conclusion section.

2. RELATED WORKS

Many strategies have been investigated for FER, they have been grouped into two main classifications: classical and deep learning-based methods. Several studies examined different FER techniques. This section provides a summary of the most modern approaches.

Zhao and Zhang [5] merged the neural network with a deep belief network (DBN) for FER. In [6] they employed a conventional CNN and two of the convolutional-pooling layers on self-collected images for face emotional.

Mollahosseini et.al. [7] investigated a more complicated architecture with four layers from type inception and two layers from convolutional-pooling. According to [8], multiple CNN learned using different filter sizes, with fully connected layers' different numbers of neurons have been tested. Although in [9] the trained of hundred CNNs have been done, they only employed a limited number in their final model and it is taken as an ensemble of CNNs. Jain et al. described the development of a CNN-RNN hybrid deep learning system in [10]. Also in [11], a hybrid system has been developed with TL, where SVM is used to classify features obtained from AlexNet. The researchers of [12] used CNN for DWT obtained features.

Liliana [13] employed architecture containing eighteen convolutional and four subsampling layers to get her results. A CNN-based clustering technique for FER has been proposed in [14]. Authors in [15] used a graph based on CNN to investigate FER from facial landmark characteristics. Experimented with various data augmentation approaches, such as using synthetic images, and discovered that a combination of synthetic data, as well as other methods, performed well for facial expression recognition [16]. Leila and Mohammad Baqer suggested extracted features as local binary features, and according to changes in points of windows, facial points get a directional motion per each facial expression. Classification is provided according to the nearest neighbor [17]

3. FACIAL EXPRESSION ANALYSIS

In human communication and behavior, facial expressions are extremely important. Paul Ekman was the first American psychologist who distinguished six basic categories of facial expressions: anger, disgust, fear, happiness, sadness, and surprise remain the same among various cultures. He then launched the Face Action Encoding System (FACS), in which he explains more than 40 different facial action units (AUs) [4].

With the advancement of artificial intelligence technologies, deep learning's remarkable progress, and the growing demand for applications that use big data, FER research will increasingly concentrate on spontaneous expressions in the wild. In such complicated situations, new solutions for problems such as multi-view, occlusion, and multi-objective are required. In general, a FER system consists of several phases, as shown in Figure 1.

4. GENERAL BACKGROUND

The FER has tested using a variety of pre-trained deep convolutional neural network models to see which one was most suited for it. This section explains CNN, as well as various DCNN models and TL.

4.1. Convolutional Neural Network (CNN)

CNN's are a class of deep learning networks that are used for processing by a grid pattern and they are most typically used to evaluate visual imagery [18] [19]. It has been built to mechanically and proactively discover spatial hierarchies of characteristics from level patterns, from low-level to high. Employing convolutional layers instead of fully connected has essentially two key benefits, the first one is sharing of parameters, and the second one is the sparsity of connections, thus reducing the number of parameters and speeding up the computation [20]. Three types of blocks (or layers) make up CNN, "convolution, pooling, and fully connected layers". The convolution and pooling layers extract features, while the full layer transfers those features to the final output. A CNN's filters slide across an image-using convolution to discover interesting patterns, and usually, its activation function is ReLU, which provides a feature map that adds to the next layer. Other layers such as "pooling, fully connected, and normalization" layers may be added after this [19].

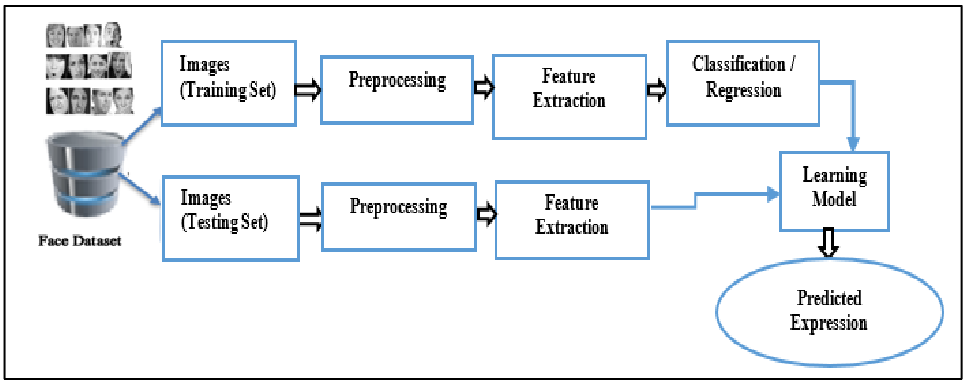


Figure 1. Conventional diagram for FER system.

A convolution layer is an important part of CNN, which is made up of stacking of arithmetic computations, as convolution which is a particular form of a linear operation. Pooling is a non-linear down-sampling technique where the pooling layer merges non-overlapping areas from one layer to generate a single value that is transferred to the next layer. Feature maps can evolve hierarchically and progressively and be more complicated as the output from one-layer feeds to the next. Training is the process of adjusting parameters like kernels (filters) to reduce the difference between outputs and ground truth labels using optimization methods such as gradient descent, backpropagation [20]. Forward propagation refers to the process of transforming input data into output data across layers. Figure 2 depicts the general architecture of a conventional CNN and training.

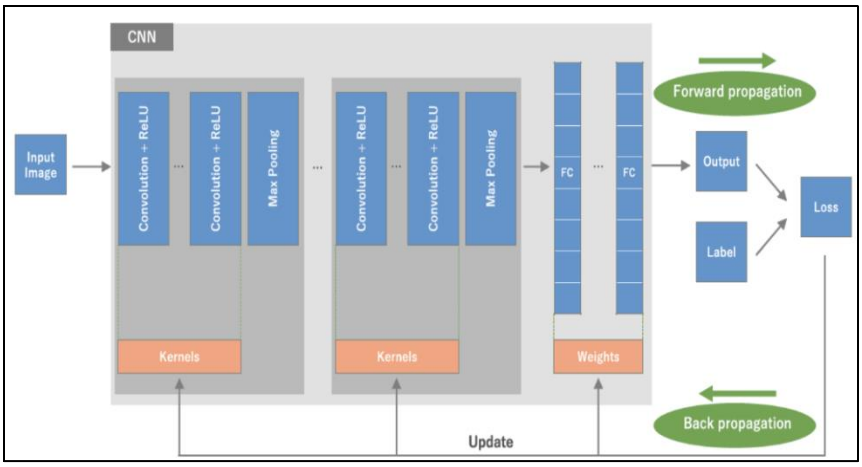


Figure 2. The architecture of a standard CNN and the training process [18].

4.2. Transfer Learning and Deep Neural Network

ImageNet is a dataset of over 15 million labeled high-resolution images belonging to roughly 22,000 categories. The images were collected from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool. Starting in 2010, as part of the Pascal Visual Object Challenge, an annual competition called the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has been held. ILSVRC uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images [22].

From 2012 to 2015, the famous CNN models that achieved the best accuracy and won in the ILSVRC were: AlexNet, where CNN employs a five-layer system [23], then ZFNet is built on the same concept, except the difference it has fewer parameters, where the large kernels swapped out for smaller ones [24], Google Net, VGG-16 [25] presented a broader model containing thirteen convolutional layers with smaller kernels.

Using sixteen convolutional layers, VGG-19 is another model from this type [26], and ResNet [27], in which the skip-connection was first proposed and it has become a key concept used by models that come following later. Currently, a few distinct ResNet models with varying depths are published, including ResNet-18, ResNet-34, ResNet-50, and ResNet -52. Later DenseNet [28] featured dense skip connections among layers in addition to a single skip link. Meaning is that every layer gets signals from the above layer, and other succeeding layers use each layer's output. The input of a layer is coupled with the channel aggregation from the previous layers. There are numerous variants of this paradigm.

DCNN features numerous hidden convolutional layers, and it works with images of large size, this makes inputs and training extremely difficult. Every DCNN model has distinct important layouts and interconnections [21]. Due to many parameters, training a large DCNN model is a difficult undertaking. A huge network frequently necessitates a significant amount of training data. Due to the high cost of data collecting and costly annotation in some disciplines, such as bioinformatics and robotics, building a large-scale well-annotated dataset is extremely challenging, limiting its evolution. However, transfer learning [29], which is focused on knowledge transfer between domains, is a potential machine learning strategy for overcoming the above difficulty, and the researchers have demonstrated that TL [30] [31] can even be highly effective in solving such issue. Deep learning has recently gotten a lot of interest from researchers as a modern categorization platform, and it has been effectively implemented in many different fields.

Transfer learning TF is a useful approach for dealing with limited training examples. Instead of learning from the beginning, the model might begin utilizing the pre-trained weights [20]. It attempts to transfer the knowledge from the source to the target domain by loosening the requirement that the training set is independently and identically distributed as the test data [32]. This will have a significant positive impact on a variety of areas that are hard to enhance due to a lack of training examples. Figure 3 depicts a conceptual diagram of the TR approach.

TL is a strategy that enables employing representations of knowledge obtained through a variety of tasks that have similar applicability. It has been observed that the TL performs better if the two tasks were identical [30]. More recently, it has been explored on tasks that are not related to its training, also it has shown to be effective [33].

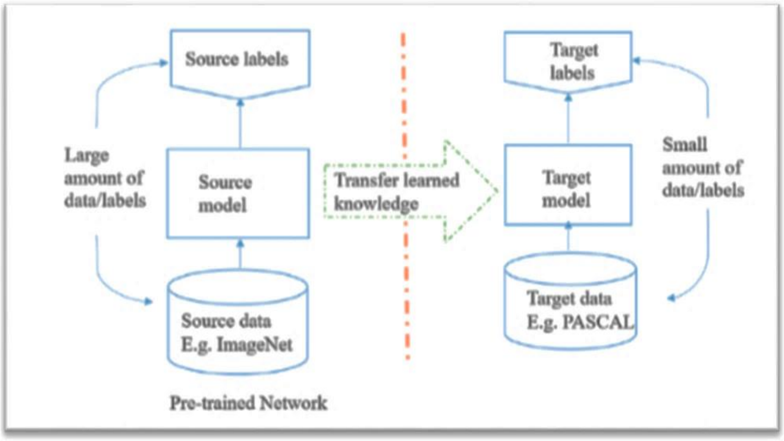

Source labels
Large amount of data/labels
Source model
Source data E.g. ImageNet
Pre-trained Network
Transfer learned knowledge
Target labels
Target model
Small amount of data/labels
Target data E.g. PASCAL

Figure 3. Conceptual diagram of TR: Pre-trained using ImageNet and retraining using PASCAL dataset [20].

5. THE PROPOSED MODEL

In this section, the proposed model for FER has been described. This model includes two parts: the features extraction part and the classifier part. The features extraction part result from integrating the features we get from pre-trained models MobileNetV2 and EfficientNetB0 that have trained on the ImageNet dataset. The training of the models will be in two cases as described later. The best one is taken as the final best model.

5.1. The Fer2013 Dataset

The dataset that has been used in this work is the FER2013 dataset, which consists of 35887 grayscale images with sizes (48x48), all images are the cropped faces, samples of these images are shown in Figure 4.

FER2013 dataset has the basic seven expressions, are "angry, disgust, fear, joy, neutral, sad, surprise". The challenges of these images are (have a small size and the low resolution), in addition to being highly imbalanced, because the number of images is different from one emotion to others, making it hard to obtain a model that can work well and accurately for all expressions, also some of the images are rotated, occluded, and with various illumination.

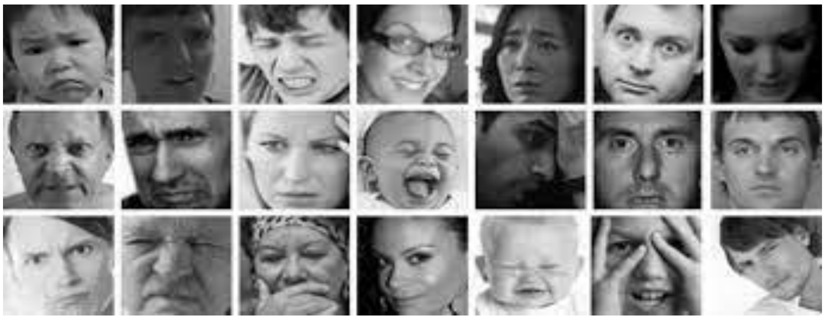


Figure 4. Samples of images for different emotions from FER2013 Dataset

5.2. Preprocessing The Dataset

The preprocessing can be summarized as a method for resizing the images to the proper size as the model requires, converting images to RGB, and reshaping the image array to desired dimensions. The images and the corresponding label will be saved as (.pkl) file; " This is a file generated by Pickle, a Python module that allows items to be serialized as files on disk and then deserialized back into the application as needed. It reloaded into memory during execution time to save space in storage or transition. The file can be used directly for reading data instead of reading the images every training, thus will reduce the required time taken for training and testing, also data augmentation has been used to increase sample size, reduce overfitting, and increase generalization. The data augmentations that have been used are (Flip, Shift, Rotate, and Zoom). In this work, the distribution of data as training and testing data are unchanged of data images as stated in the original dataset page, with 28709 training samples and 3589 validation samples, without adding any external images as the other researchers did, where a lot of researchers used external images from other datasets or those that are self- obtained images.

5.3. The Architecture Of The Proposed Model

The proposed model is structured as shown in Figure 5, the weights from the pre-trained models such as (EfficientNetB0 and MobileNetV2) transferred to the newly built model by taking only the features extraction part from it, leaving its old classifier, and replaced with new classifier. Features of EfficientNetB0 and MobileNetV2 models have been combined to create a feature extractor part for the model.

The Classifier part is created by adding two Dense-Dropout layers with (4096 and 1024) neurons respectively. The activation function that been used in these layers was of type ReLU. The last layer of the model is the Dense layer with seven neurons (because the model will be used to "classify image to one of seven expressions") and the activation function was Softmax.

5.4. Training The Models

In general, when the models have been trained using transfer learning, the training speed can be faster, then the weights of its feature extraction layers (ignore the classification part) can be used to initialize the feature extraction layers of a new model. In addition, to increase the accuracy and generalization of the model, fine-tuning techniques are used. The proposed hybrid model described in the previous subsection has been training using features extraction parts pre-trained on Imagenet and separately on Fer2013 Itself. Also, for single and hybrid models ,the training is done using two cases as mentioned bellow and the best obtained results for the models are listed in this paper.

o Case 1: freezing the features extraction part and training only the classification part, the extracted general features will be good enough at the start. Freezing the extraction part allows the classification layers to get some knowledge about what features to look for in the output, it takes less training time than other cases because the parameters that need training are less.

o Case 2: fine-tune the extraction part and the classification part, so the whole model is trained without freezing any part. This may make the learning process give better accuracy, but it takes more training time.

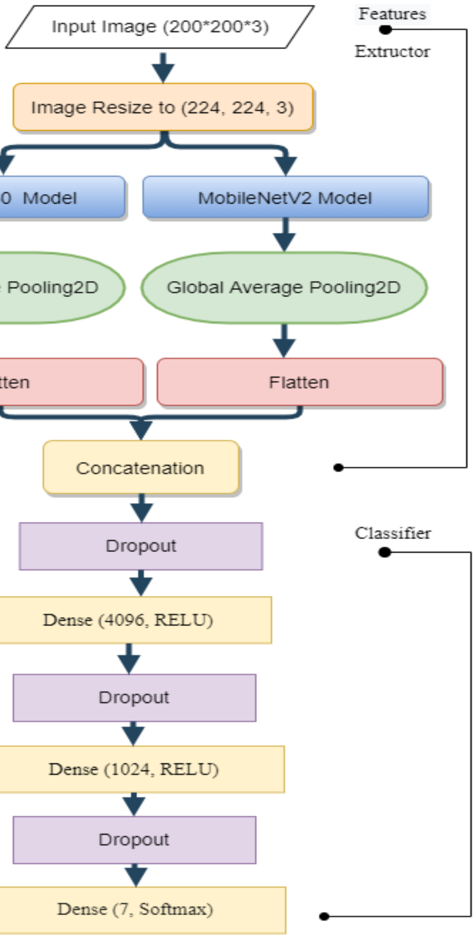


Figure 5. The structure of the proposed hybrid model

6. THE EXPERIMENTAL RESULTS

The parameters that used for training all models are listed in Table 1, the following subsections explain the results obtained from training two pre-trained models separately, and that get from the proposed combining models trained in two cases (directly using the weights from pre-trained models on ImageNet and by using the two pre-trained models weights trained on FER-2013 separately).

Table 1. Training parameters

Parameter	Value
Input Image Size	200x200
Image Color Space	RGB 3 Layers
Batch Size	32
No of Epochs	100
Dropout Rate	0.2
optimization function	ADAM
Learning Rate	0.001

6.1. The EfficientnetB0 Model

In this model, the weights from the EfficientNetB0 model are transferred to the newly built model by replacing its old classifier with the new classifier consisting of two Dense-Dropout layers with (4096 and 1024) neurons respectively, and Rectified Linear Unit (ReLU) activation function. Also, the last layer is the Dense layer with 7 neurons and Softmax activation function.

The structure of this model is shown in Figure 6.The model has been trained by fine-tuning both parts (features extractor and the classifier). Model accuracy and Loss for this case are plotted in Figure 7 and the best accuracy was (73.32) at epoch 95. The confusion matrix and the classification report on testing the final best model on a test set of FER2013 are shown in Figure 8.

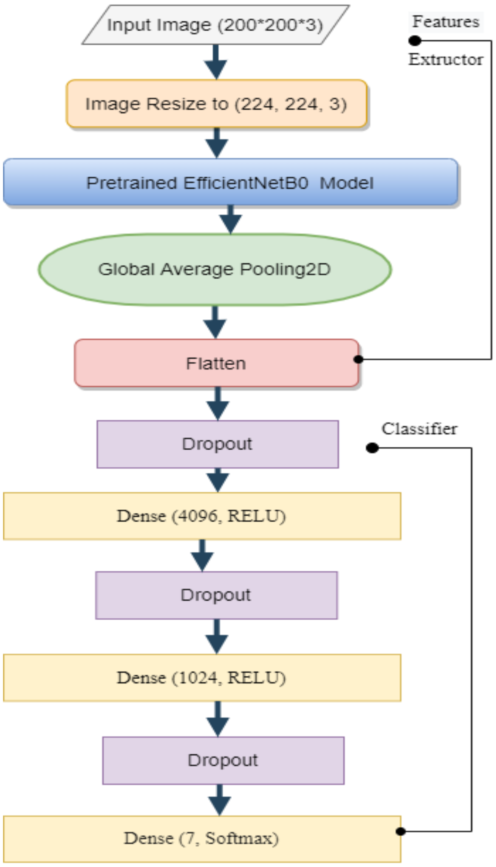


Figure 6. The structure of the proposed EfficientNetB0 model.

6.2. The MobilenetV2 Model

The MobileNetV2 has been trained using the same strategy for EfficientNetB0 in the previous subsection with the same structure shown in Fig. 6 except that EfficientNetB0 was replaced by MobileNetV2. The best accuracy for the training with fine-tuning was (0.71038) at epoch 85. Model accuracy and Loss have been plotted in Figure 9. The confusion matrix and the classification report on testing the final best model on a test set of FER2013 are shown in Figure 10.

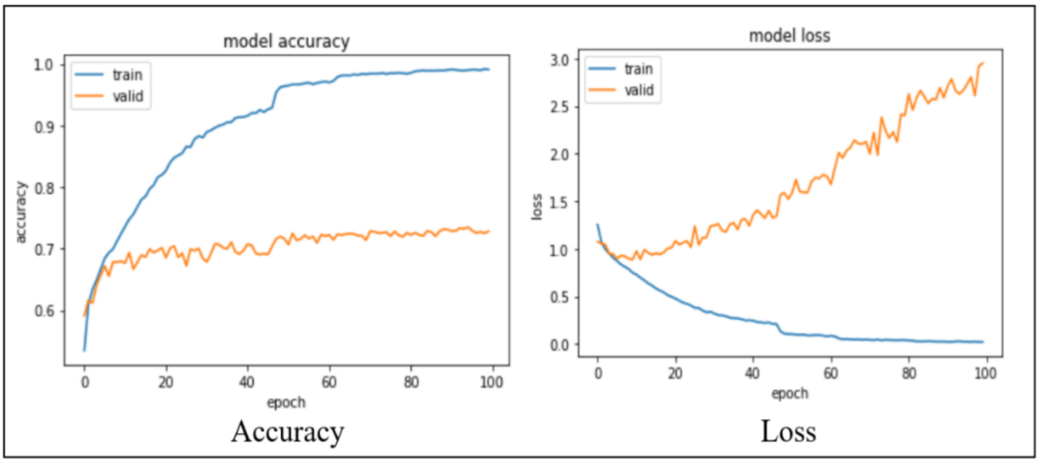
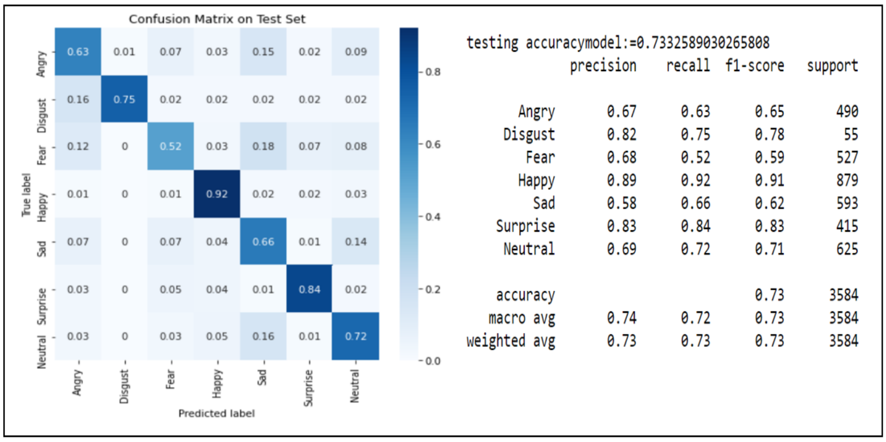


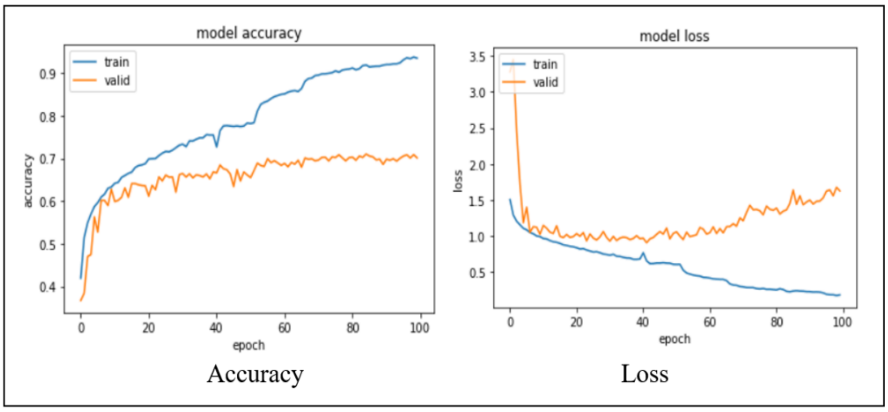
Figure 7. The proposed EfficientNetB0 Model: accuracy and loss after fine-tuning



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.63	0.01	0.07	0.03	0.15	0.02	0.09
Disgust	0.16	0.75	0.02	0.02	0.02	0.02	0.02
Fear	0.12	0	0.52	0.03	0.18	0.07	0.08
Happy	0.01	0	0.01	0.92	0.02	0.02	0.03
Sad	0.07	0	0.07	0.04	0.66	0.01	0.14
Surprise	0.03	0	0.05	0.04	0.01	0.84	0.02
Neutral	0.03	0	0.03	0.05	0.16	0.01	0.72

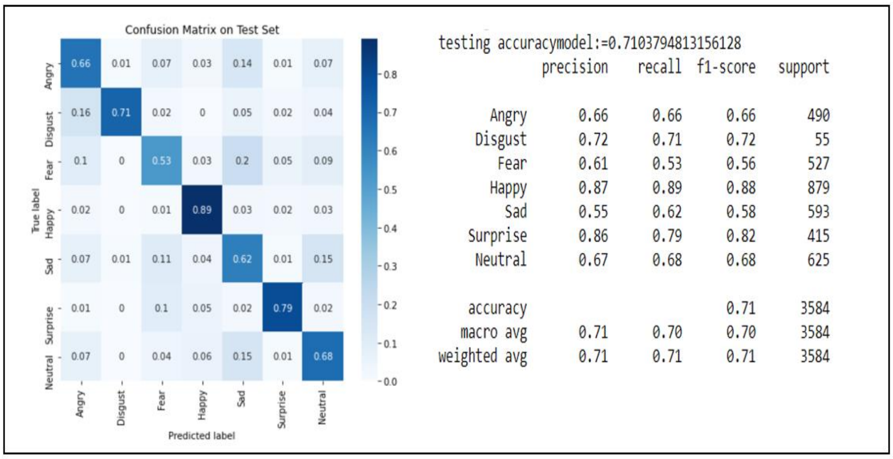
```
testing accuracymodel:=0.7332589030265808
              precision    recall  f1-score   support

       Angry       0.67      0.63      0.65       490
     Disgust       0.82      0.75      0.78        55
        Fear       0.68      0.52      0.59       527
       Happy       0.89      0.92      0.91       879
         Sad       0.58      0.66      0.62       593
    Surprise       0.83      0.84      0.83       415
     Neutral       0.69      0.72      0.71       625

    accuracy                           0.73      3584
   macro avg       0.74      0.72      0.73      3584
weighted avg       0.73      0.73      0.73      3584
```

Figure 8. The confusion matrix and the classification report of the final best EfficientNetB0 model on FER2013.



Accuracy Loss

Figure 9. The proposed MobileNetV2 model: accuracy and loss after fine-tuning



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.66	0.01	0.07	0.03	0.14	0.01	0.07
Disgust	0.16	0.71	0.02	0	0.05	0.02	0.04
Fear	0.1	0	0.53	0.03	0.2	0.05	0.09
Happy	0.02	0	0.01	0.89	0.03	0.02	0.03
Sad	0.07	0.01	0.11	0.04	0.62	0.01	0.15
Surprise	0.01	0	0.1	0.05	0.02	0.79	0.02
Neutral	0.07	0	0.04	0.06	0.15	0.01	0.68

```
testing accuracymodel:=0.7103794813156128
              precision    recall  f1-score   support

       Angry       0.66      0.66      0.66       490
     Disgust       0.72      0.71      0.72        55
        Fear       0.61      0.53      0.56       527
       Happy       0.87      0.89      0.88       879
         Sad       0.55      0.62      0.58       593
    Surprise       0.86      0.79      0.82       415
     Neutral       0.67      0.68      0.68       625

    accuracy                           0.71      3584
   macro avg       0.71      0.70      0.70      3584
weighted avg       0.71      0.71      0.71      3584
```

Figure 10. The confusion matrix and the classification report of testing the final best MobileNetV2 model on the FER2013 test set

6.3. The Hybrid Model Using Features Extraction Parts Pre-Trained On Imagenet.

The proposed hybrid model described in the previous section has been trained in the two cases with the following results:

Case 1: training by freezing the features extraction parts of the model, the accuracy and loss have been plotted in Figure 11; the best accuracy was 0.588 at epoch 147.

Case 2: training by fine-tuning (unfreezing) the features extraction parts of the model, the accuracy and Loss have plotted in Figure 12, in this case, the best accuracy was 0.7322 at epoch 74. A confusion matrix and a snapshot of the classification report contain "precision, recall, F1-score, support" metrics obtained from testing the hybrid model, in this case, are shown in Figure 13.

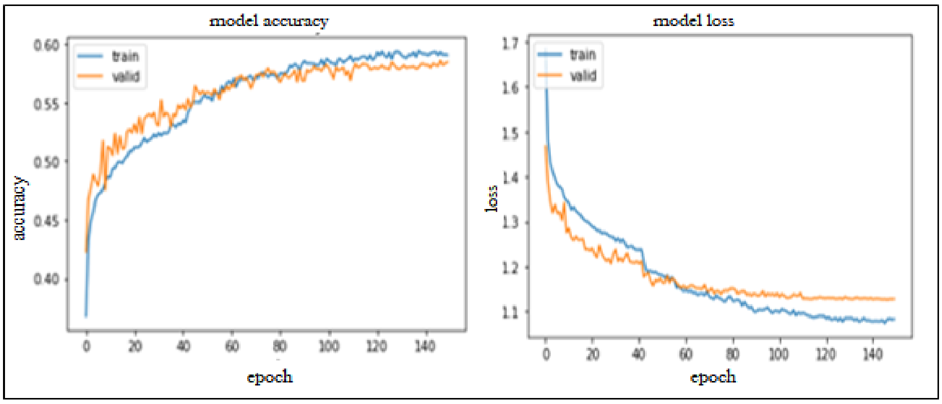


Figure 11. The proposed hybrid model using pre-trained features extraction parts pre-trained on ImageNet: accuracy and loss after case 1

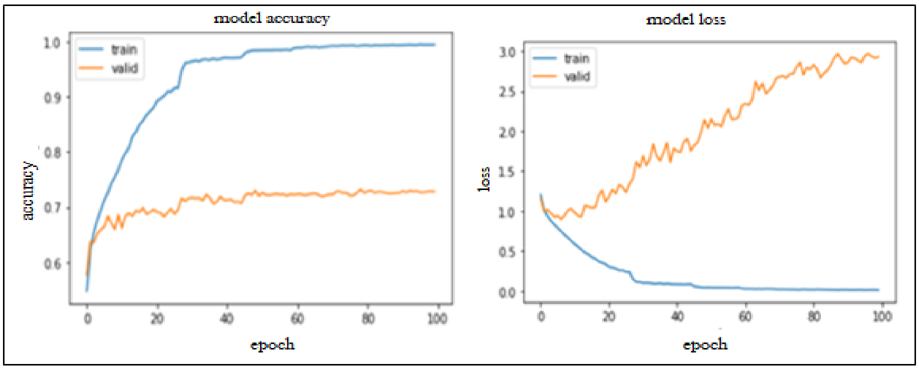
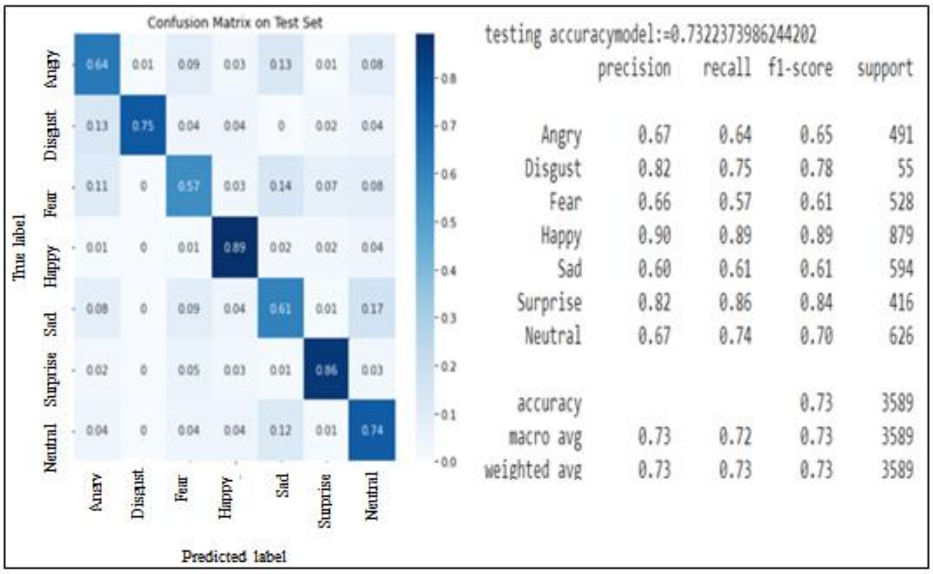


Figure 12. The proposed hybrid model using pre-trained features extraction parts pre-trained on ImageNet: accuracy and loss after case 2



```
testing accuracymodel:=0.7322373986244202
              precision    recall  f1-score   support

       Angry       0.67      0.64      0.65       491
     Disgust       0.82      0.75      0.78        55
        Fear       0.66      0.57      0.61       528
       Happy       0.90      0.89      0.89       879
         Sad       0.60      0.61      0.61       594
    Surprise       0.82      0.86      0.84       416
     Neutral       0.67      0.74      0.70       626

    accuracy                           0.73      3589
   macro avg       0.73      0.72      0.73      3589
weighted avg       0.73      0.73      0.73      3589
```

Figure 13. The confusion matrix and the classification report of testing the proposed hybrid model (pre-trained on ImageNet) after case 2

6.4. The Hybrid Model Using Features Extraction Parts Pre-Trained Separately On Fer2013 Itself.

The weights from EfficientNetB0 and MobileNetV2 models that trained speratly as described in subsection 6.1 and 6.2 are used as features extraction parts for the proposed hybrid model. By freezing these features extraction parts, the model got a higher accuracy up to (0.7439) at epoch number 16. The increasing accuracy results may be because of the good features (weights) obtained from the feature extraction parts. The accuracy and Loss for the final best model are shown in Figure 14. In addition, the confusion matrix and a snapshot of the classification report contain "precision, recall, F1-score, support" metrics obtained from testing the best hybrid model are shown in Figure 15.

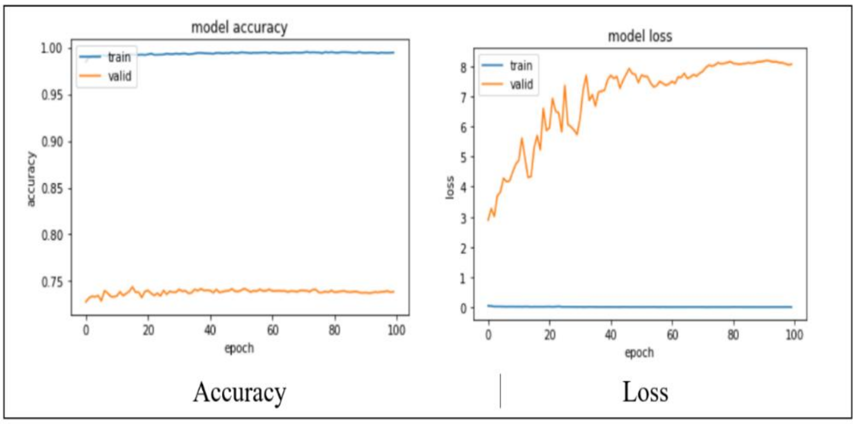
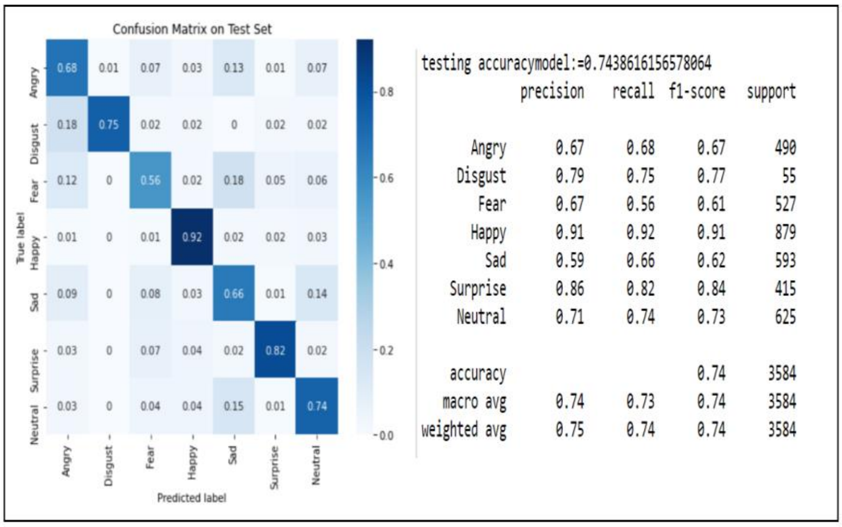


Figure 14. The proposed hybrid model using features extraction parts pre-trained separately on FER2013 itself: accuracy and loss



	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	0.68	0.01	0.07	0.03	0.13	0.01	0.07
Disgust	0.18	0.75	0.02	0.02	0	0.02	0.02
Fear	0.12	0	0.56	0.02	0.18	0.05	0.06
Happy	0.01	0	0.01	0.92	0.02	0.02	0.03
Sad	0.09	0	0.08	0.03	0.66	0.01	0.14
Surprise	0.03	0	0.07	0.04	0.02	0.82	0.02
Neutral	0.03	0	0.04	0.04	0.15	0.01	0.74

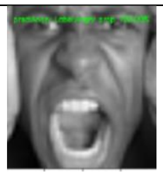
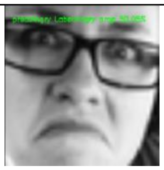
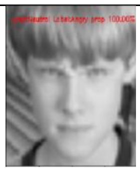
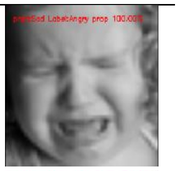
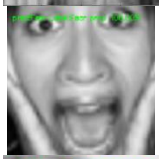
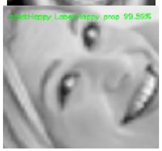
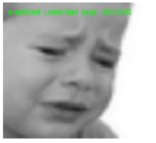
```
testing accuracymodel:=0.7438616156578064
              precision    recall  f1-score   support

       Angry       0.67      0.68      0.67       490
     Disgust       0.79      0.75      0.77        55
        Fear       0.67      0.56      0.61       527
       Happy       0.91      0.92      0.91       879
         Sad       0.59      0.66      0.62       593
    Surprise       0.86      0.82      0.84       415
     Neutral       0.71      0.74      0.73       625

    accuracy                           0.74      3584
   macro avg       0.74      0.73      0.74      3584
weighted avg       0.75      0.74      0.74      3584
```

Figure 15. The confusion matrix and the classification report of testing the final best hybrid model (features extraction parts pre-trained separately) on the FER2013 test set.

7. TESTING THE BEST MODEL

Table 2 presents samples of images on which the best hybrid model was tested to distinguish its expression, the first three columns containing samples of correctly predicted images which are labeled with green text, while the last three columns indicate samples of incorrectly predicted images which they labeled with red text. As noticed, when looking for images which are incorrectly recognized by the system, for example, images in the last column in the rows of the expressions “Fear, Sad, Natural and Surprise”, it is clear that most of these images have happiness expression, but it is incorrectly labeled in the FER2013 dataset. This means that the suggested model predicted the correct expression in spite that it is regarded as wrong predictions from the model. There are a lot of incorrectly labeled images in the fer2013 dataset which affects a lot in the accuracy but the suggested model correctly recognized them. Examples of these images are shown in the last two columns of Table 2.

Table 2. Examples of correctly and incorrectly predicted images from testing the best model

True label	True prediction			False prediction		
	1	2	3	1	2	3
Angry						
Disgust						
Fear						
Happy						
Sad						
Surprise						
Neutral						

8. PERFORMANCE COMPARISON

Table 3 listed all the results of training the models that have been done. Training the model from the base means that it is depending on the pre-trained weights from the ImageNet dataset. Training it using loading weights means that its weights are obtained from training each model (EfficientNetB0, MobileNetV2) separately on the FER2013 dataset. FER2013's prior reported classification accuracies are shown in Table 4 with a comparison with the best of our models.

The best-reported accuracy that was achieved without using any additional data was 73.28 %. In this work we did not add any data for training the models and depending only on the dataset with its all challenges, the best-achieved accuracy was 74.39 % by combining the two models (EfficientNetB0, MobileNetV2) that are trained separately on FER2013.

Table 3. The Comparison between our models.

The model	Accuracy %	Best Epoch
EfficientNetB0	73.32	95
MobileNetV2	71.04	85
Hybrid Model Using Features Extraction Parts Pre-Trained on Imagenet : FF[a]	58.63	147
Hybrid Model Using Features Extraction Parts Pre-Trained on Imagenet. : TT[b]	73.22	74
Hybrid Model Using Features Extraction Parts Pre-Trained on Fer2013 : FF[a]	74.39	16

a: (Mobilenetv2_classifier.trainable = False, EfficientNetB0.trainable = False)

b:(MobileNetV2_classifier.trainable = True and EfficientNetB0.trainable = True)

Table 4. The Comparison Comparison with recent models on the FER2013 dataset.

Year	The model/ Method	Accuracy %
Proposed	Hybrid Model by freezing the Features Extraction Parts Pre-Trained on Fer2013.	**74.39**
2020	Residual Masking Network (using extra data) [35]	74.14
Proposed	EfficientNetB0 model with Fine tuining	73.32
2021	VGGNET [34]	73.28
Proposed	Hybrid Model by Fine tuining the Features Extraction Parts Pre-Trained on Imagenet	73.22
2016	VGG [36]	72.70
2016	Resnet [36]	72.40
2021	CNN Hyperparameter Optimisation [37]	72.16
2016	Inception [36]	71.60
2021	ARM(ResNet-18) [38]	71.38
Proposed	MobileNetV2 model with Fine tuining	71.04
2019	AttentionalConvNet [39]	70.02
2016	Con+Inception layer [7]	66.40
2019	VGG+SVM [40]	66.31

9. DISCUSSION

This research provides a method for recognizing facial emotions using pre-trained deep Neural Networks and the TL by developing single and hybrid models. The pre-trained models used in this work are EfficientNetB0 and MobileNetV2 models, the single pre-trained models are trained by using a new classifier and fine-tuning the features extraction part.the proposed hybrid model contains a mix of two single pre-trained models which are EfficientNetB0 and MobileNetV2 models to categorize facial expression into seven categories. Also, the hybrid model uses the same classifier which used with single models.

When examining Table 3 ,It is noticed that fine-tuning the feature extraction part of the single models gives better accuracy compared with cases that the feature extraction part is freezing. This is because the FER2013 dataset that contains only faces is very different from the Imagenet dataset that single models trained with, Thus, the features extracted by using these models differ greatly from the facial features. Especially in the final layers that give a clearer perception of the image.For the hybrid model, the results are different depending on the data set on which the individual models were previously trained and if the training is with or without fine-tuning.

Due to the limitation of the resources like (RAM, GPU, CPU), the training was for just 100 epochs, except that without tuning there is trying to train for 150 epochs to get better accuracy, but as we see there is no enhancement. For more epochs, this will slow the training and elapse more time because training the model on a large dataset required a lot of computational operations and memory.

As the results have shown, our models outperform other state-of-the-art classification methods without using any additional data and the best model gives accuracy equal to 74.39 %. This accuracy is highest than all the models that are not used extra data (the highest accuracy for methods that do not use extra data was 73.28 [34]).

It even outperforms the method that is considered the second-highest accuracy for the classification of expressions [35] using the FER2013 dataset with 74.14 % accuracy according to the site (https://paperswithcode.com/sota/facial-expression-recognition-on-FER-2013). In addition, their model is more complex and requires high resources.In this case, the best-proposed model in this paper ranked the second position among all methods in the classification of facial expressions for the FER-2013 database, exceeding all methods that do not use extra data; also, it can be used in devices with limitations resources.

The results show that the best accuracy of the suggested models is 74.39% for the hybrid model, and 73.33% for Fine-tuned the single EfficientNetB0 model, while the highest accuracy for previous methods was 73.28%[34]. Thus, the hybrid and single models outperform other state of art classification methods without using any additional, the hybrid and single models ranked in the first and second position among these methods. Also, The hybrid model has even outperformed the second-highest accuracy for the classification of expressions [35] which used extra data with 74.14 % accuracy according to the site (https://paperswithcode.com/sota/facial-expression-recognition-on-FER-2013). Also, The incorrectly labeled images in the dataset unfairly reduce accuracy but our best model recognized their actual classes correctly.

The accuracy of the happy and surprised expressions was the highest, but the fear, sadness, and angry expressions had less accurate. It has maybe natural because the first two expressions have more images than the rest and their features are distinctive. On the other side, even for a human, it is difficult for classifying some of such images. This may be because the features for these expressions are more similar to each other, which leads to confusion and inaccurate classification.

Table 5 displays the largest amount of confusion between expressions. The largest confusion was between Disgust expressions and angry expressions where the model predicted 0.18 % of Disgust expression as angry expression, sad expression confused with neutral expression with value 0.15, fear expression with sad expression by 0.14.

Furthermore, it is clear that only anger expression has been predicated as disgust expression, and in a very small percentage that is almost negligible about 0.01. Thus the model does not predict wrongly any other expressions as disgust.

Table 5. The Most Confusion between Expressions.

True Expression	Confused/ Predicted Expression	Value
Angry	Sad	0.13
Disgust	Angry	0.18
Fear	Sad	0.18
Happy	Neutral	0.03
Sad	Neutral	0.14
Surprise	Fear	0.07
Neutral	Sad	0.15

10. CONCLUSION

In this thesis, a system for facial expression recognition from images is implemented by using different pre-trained models and the proposed hybrid model. two pre-trained models have been applied after changing classifiers then constructing the hybrid model. The proposed hybrid model consists of merging two models and training them both to rely on the ImageNet data set and by using weights from each model pre-trained on the Fer2013 data set separately.However, facial expression recognition is not an easy task and still suffers from difficulties especially since some facial expressions are almost similar and the confusion happens between them for the same person or with another person. In addition, the available datasets may have many conditions in various environments, like there was various lighting, occlusion for some facial features, etc., which adds difficulties in classification. The proposed model Has the ability to address most of these problems found in the images and is not limited to images of the front faces only. A list of conclusions from this paper are listed below:

- If the database is completely or significantly different from the one on which the model was previously trained, the performance will not be very effective. Because only elementary features such as basic lines and shapes are extracted correctly, but more complex features will not be extracted well using previously trained weights.

- If the model is previously trained on a database similar or close to the one on which training is required, in this case, the results will be better and the performance more effective because the weights are set correctly to extract almost all the features, even the complex ones.

- The proposed hybrid model outperforms other methods without using any additional images, whereas more approaches used additional data to achieve more accuracy.

- Although the best models performed differently, they all produced encouraging results, exceeding various state-of-the-art techniques for the classification of facial expressions for the FER-2013 database

- There are a lot of incorrectly labeled images in the fer2013 dataset which unfairly reduces the accuracy but the model correctly recognized them.

Future work could include training the model with various values for training parameters (batch size, training epochs, optimization functions, and so on), also can train the model for other cases like freezing one of its feature extraction parts and fine-tuning the others..

REFERENCES

[1] Liew, C.F., Yairi, T., Facial expression recognition and analysis: A comparison study of feature descriptors, IPSJ Trans. Comput. Vis. Appl., 2015, 7, 104–120. DOI:10.2197/ipsjtcva.7.104.

[2] Ko, B.C., A Brief Review of Facial Emotion Recognition Based, Sensors, 2018 Jan 30;18(2):401. DOI: 10.3390/s18020401.

[3] Huang, Y., Chen, F., Lv, S., Wang, X., symmetry Facial Expression Recognition : A Survey, *Symmetry* 2019, *11*(10), 1189; DOI: 10.3390/sym11101189

[4] Li, S., Deng, W., Deep Facial Expression Recognition: A Survey, IEEE Trans. Affect. Comput. 2903 (2020). DOI:10.1109/TAFFC.2020.2981446.

[5] Zhao, X., Shi, X., Zhang, S., Facial expression recognition via deep learning, IETE Tech. Rev. (Institution Electron. Telecommun. Eng. India). 32, 347–355 (2015). DOI:10.1080/02564602.2015.1017542.

[6] Pranav, E.; Kamal, S.; Chandran, C.S.; Supriya, M., Facial emotion recognition using convolutional neural network, In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)., 2020 pp. 317–320. DOI:10.1109/ICIEM51511.2021.9445346.

[7] Mollahosseini, A., Chan, D., Mahoor, M.H., Going deeper in facial expression recognition using deep neural networks, 2016 IEEE Winter Conf. Appl. Comput. Vision, WACV 2016. (2016). DOI:10.1109/WACV.2016.7477450.

[8] Pons, G., Masip, D., Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis, IEEE Trans. Affect. Comput. 9, 343–350 (2018). DOI:10.1109/TAFFC.2017.2753235.

[9] Wen, G., Hou, Z., Li, H., Li, D., Jiang, L., Xun, E., Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition, Cognit. Comput., 2017, 9, 597–610. DOI:10.1007/s12559-017-9472-6.

[10] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M. US CR. Pattern Recognit. Lett. (2018). DOI:10.1016/j.patrec.2018.04.010.

[11] Shaees, S., Naeem, H., Arslan, M., Naeem, M.R., Ali, S.H., Aldabbas, H., Facial emotion recognition using transfer learning, In 2020 International Conference on Computing and Information Technology (ICCIT-1441). pp. 1–5. IEEE (2020).

[12] Bendjillali, R.I., Beladgham, M., Merit, K., Improved Facial Expression Recognition Based on DWT Feature for Deep CNN, (2019). DOI:10.3390/electronics8030324.

[13] Liliana, D.Y., Emotion recognition from a facial expression using deep convolutional neural network, J. Phys. Conf. Ser. 1193, (2019). DOI:10.1088/1742-6596/1193/1/012004.

[14] Shi, M., Xu, L., Chen, X., A Novel Facial Expression Intelligent Recognition Method Using Improved Convolutional Neural Network, IEEE Access. 8, 57606–57614 (2020). DOI:10.1109/ACCESS.2020.2982286.

[15] Ngoc, Q.T., Lee, S., Song, B.C., Facial landmark-based emotion recognition via directed graph neural network, Electron. 9, (2020). DOI:10.3390/electronics9050764.

[16] Porcu, S., Floris, A., Atzori, L., Evaluation of data augmentation techniques for facial expression recognition systems. Electron. 9, 1–12 (2020). DOI:10.3390/electronics9111892.

[17] Sahu, M., Dash, R., "A survey on deep learning: Convolution neural network (CNN). Springer Singapore (2021). DOI:10.1007/978-981-15-6202-0_32.

[18] Leila Farmohammadi, Mohammad Baqer Menhaj, Facial Expression Recognition Based on Facial Motion Patterns, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 2015, Vol. 3, No. 4, pp. 177~184

[19] Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K., Convolutional neural networks: an overview and application in radiology. Insights Imaging. 9, 611–629 (2018). DOI:10.1007/s13244-018-0639-9.

[20] Wu, J., Introduction to Convolutional Neural Networks. Introd. to Convolutional Neural Networks. 1–31 (2017).

[21] Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A.S., Asari, V.K., A state-of-the-art survey on deep learning theory and architectures. Electron. 8, (2019). DOI:10.3390/electronics8030292.

[22] Russakovsky O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 2015. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[23] Krizhevsky A., Sutskever I., Hinton G., ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012, 1106-1114

[24] Zeiler, M.D., Fergus, R., Visualizing and understanding convolutional networks, Lect. Notes Comput. Sci. 2014. 8689 LNCS, 818–833. DOI:10.1007/978-3-319-10590-1_53.

[25] Simonyan, K., Zisserman, A., Very deep convolutional networks for large-scale image recognition, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 1–14 (2015).

[26] Nair V., Hinton, G.E., Rectified linear units improve restricted Boltzmann machines, ICML 2010 - Proceedings, 27th Int. Conf. Mach. Learn. 807–814 (2010).

[27] He, K., Zhang, X., Ren, S., Sun, J., Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 770–778 (2016). DOI:10.1109/CVPR.2016.90.

[28] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. , Densely connected convolutional networks, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-January, 2261–2269 (2017). DOI:10.1109/CVPR.2017.243.

[29] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., A Comprehensive Survey on Transfer Learning, Proc. IEEE. 109, 43–76 (2021). DOI:10.1109/JPROC.2020.3004555.

[30] Oquab, M., Bottou, L., Laptev, I., Sivic, J., Learning and transferring mid-level image representations using convolutional neural networks, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 1717–1724 (2014). DOI:10.1109/CVPR.2014.222.

[31] Yosinski, J., Clune, J., Bengio, Y., Lipson, H., How transferable are features in deep neural networks? , Adv. Neural Inf. Process. Syst. 2014. 4, 3320–3328.

[32] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., A survey on deep transfer learning, Lect. Notes Comput. Sci. 2018. 11141 LNCS, 270–279. DOI:10.1007/978-3-030-01424-7_27.

[33] Bukar, A.M., Ugail, H., Automatic age estimation from a facial profile view, IET Comput. Vis. 2017. 11, 650–655. DOI:10.1049/iet-cvi.2016.0486.

[34] Khaireddin, Y., Chen, Z., Facial Emotion Recognition : State of the Art Performance on FER2013,

[35] Pham, L., Vu, T.H., Tran, T.A.: Facial Expression Recognition Using Residual Masking Network, In 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4513–4519. IEEE (2021).

[36] Pramerdorfer, C., Kampel, M., Facial Expression Recognition using Convolutional Neural Networks: State of the Art, (2016).

[37] Vulpe-Grigoraşi, A., Grigore, O., Convolutional Neural Network Hyperparameters optimization for Facial Emotion Recognition, In 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE). 2021. pp. 1–5. IEEE 2021.

[38] Shi, J., Zhu, S., Learning to amend facial expression representation via de-albino and affinity, arXiv Prepr. arXiv2103.10189. (2021).

[39] Minaee, S., Minaei, M., Abdolrashidi, A., Deep-emotion: Facial expression recognition using an attentional convolutional network, Sensors. 2021. 21, 3046.

[40] Georgescu, M.-I., Ionescu, R.T., Popescu, M., Local learning with deep and handcrafted features for facial expression recognition, IEEE Access. 2019. 7, 64827–64836.