Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

DLFER: Deep Learning-based Cascaded Approach for Facial Emotion Recognition

Payman Hussein Hussan [1*]

[1] Al-Furat Al-Awsat Technical University, Babylon Technical Institute, Department of Computer Networks and Software Techniques, Babil, 51015, Iraq, E-mail: inb.beman10@atu.edu.iq

[*]Corresponding author E-mail: inb.beman10@atu.edu.iq

https://doi.org/10.46649/fjiece.v4.1.19a.25.3.2025

Abstract. *Facial expression identification has garnered considerable attention in recent years owing to its extensive applicability across various domains, including human-computer interaction, market research, and healthcare. The primary objective of Facial Emotion Recognition (FER) is to correlate various facial expressions with their corresponding emotional states. Advancements in deep learning have significantly enhanced the recognition accuracy of FER technology relative to conventional approaches. This study seeks to improve the accuracy for facial emotion identification by introducing a Deep Learning Cascaded Network (DLFER) founded on the EF-FER1 and EF-FER2 architectures. The hyperparameters of the EfficientNet-B3 network were fine-tuned to improve facial expression representation and classification. The trials utilized the widely recognized Facial Expression Recognition 2013 (FER2013) dataset, comprising 35,887 greyscale photos of faces, each linked to one of seven specific emotions. The model's performance was assessed using accuracy, precision, recall, and F1 score. A comparative analysis was also performed with similar modern studies. The trials demonstrated that the Cascaded Network (DLFER) attained a classification accuracy of 82.09%, surpassing that of state-of-the-art models.*

Keywords:.*Deep learning ; Facial emotion recognition (FER) ; transfer learing ; EfficientNet-B3; Fine-tuning model*

1. INTRODUCTION

Emotions are essential human characteristics that play important roles in social communication. Humans express emotions through several means, including facial expressions, verbal communication, and body language. Facial expression analysis is the most prominent and extensively studied aspect of emotion recognition [1].

Humans employ diverse communication strategies to convey their messages. One of them is expressing their thoughts, emotions, and attitudes through facial expressions. An individual's emotional condition can be encapsulated by facial expressions resulting from a combination of nuanced facial movements. Eckman devised the Facial Action Coding System (FACS) to categorize facial movements. FACS was initially utilized to assess and analyze the activity of all minor face muscles. Facial expression encompasses alterations in the muscles, eyes, and head. His research posited seven universal emotions applicable to all individuals: happiness, sadness, surprise, fear, rage, disgust, and contempt. Human-computer interaction is crucial in deep learning models for the recognition of emotions via facial expressions [2].

Facial emotion recognition is a technology that identifies and analyzes human emotions using facial expressions. It is an interdisciplinary domain that integrates computer vision and machine learning. The identification of facial emotions seeks to create algorithms and systems capable of effectively recognizing

and interpreting an individual's emotional state through the analysis of facial features. Human emotions are conveyed through diverse facial expressions, including joy, sadness, anger, fear, surprise, and disrespect. These expressions necessitate the involvement of various facial muscles, which can be documented and examined through photographs or video recordings. Facial emotion detection systems utilize this data to identify pertinent traits and patterns associated with particular emotions [3].

The advancement of deep learning algorithms led to a significant breakthrough in the domain of emotion recognition from facial photographs. Deep learning approaches have demonstrated exceptional efficacy in emotion detection using face features by emulating the architecture and operation of the human brain. Furthermore, deep learning systems can discern intricate patterns, extract features from extensive datasets, and generalize their learning abilities to novel data. The identification of emotions using facial characteristics utilizing deep learning algorithms has become a promising field of research[4].

Facial expressions serve primarily as indicators of emotions and emotional information. Contemporary researchers employ deep CNNs to integrate feature extraction and emotion categorization into a singular procedure. Talegaonkar el al. [5] introduced a deep learning model utilising CNNs for real time emotion classification via webcam. The model was developed to identify the user's emotions when viewing movie trailers or listening to video lectures. The model was assessed using the FER2013 dataset and attained an accuracy of 60.12%. Minaei and Abdolrashidi in [6]employed a deep learning method that focused on essential facial traits and surpassed previous models across various datasets. The scientists utilised a modelling method informed by the classifier's results to identify critical face regions associated with different emotions. The model was trained on 28709 pictures, and its accuracy was reported. They attained a remarkably high accuracy rate on the test set. The weakness of the study is the proposed method's inadequacy in addressing the imbalanced nature of different emotion groups within the FER2013 dataset.

Zhu et al. [7]examined the correlation between emotional facial recognition and behavioral characteristics. Based on this premise, a facial emotion detection model is developed by augmenting the CNN' layers and integrating it with diverse neural networks for facial emotion recognition and achieved an accuracy of 70.2%.

In [8]Punurietal introduced a novel strategy based on the Transfer Learning methodology, termed "EfficientNet XGBoost" and achieved an accuracy of 72.5%. This model combines the strengths of these algorithms. The writers demonstrated its advantage over the novelty of the technique.

The research [9] examined facial expression recognition (FER) employing ensemble techniques that integrate pre-trained models AlexNet and InceptionV3 utilizing the FER2013 dataset. Transfer learning facilitates the utilization of pre-trained models to address data constraints. The ensemble technique attains a maximum accuracy of 73.56% following data augmentation.

Motivation

Although previous studies have shown promising findings for facial emotion recognition (FER), there is still room for improvement. This innovation would aid practitioners in various applications, including facial authentication systems, entertainment, and deepfake detection.

Therefore, to develop a high-performance approach to recognize emotions based on facial features in human faces, the following questions are addressed: Which deep learning architecture is considered the most effective for recognizing emotions in human faces? How can training strategies improve the model's performance without relying on large-scale datasets? Is there a specific approach that addresses this, and how can it contribute to enhancing the accuracy of emotion recognition? Are there specific applications or domains where facial emotion recognition is particularly relevant?

This research intends to address the questions above, which are significant for recognizing emotions based on facial features. The main contributions that this paper brings out are:

1- Proposing the Cascaded approach (DLFER), which consists of two deep learning models, EF-FER1 and EF-FER2, to enhance the results of facial emotion recognition.
2- To improve accuracy without a significant increase in computational burden, An architecture that has

learned knowledge from other domains has been used to obtain an efficient approach to facial expression recognition.

3- To extract fine-grained features from facial images, an architecture (EFFER) model had been proposed inspired by EfficientNetB3, making the model adept at recognizing subtle or nuanced emotions.

The subsequent sections of the paper are structured in the following manner: Section 2 outlines the technique utilized for the study. Section 3 presents the findings and performs an extensive series of comparative evaluations with the current accomplishments in the literature, followed by a detailed discussion and analysis of the outcomes in Section 4. Ultimately, conclusions are presented in Section 5.

2. MATERIALS AND METHODS

This section delineates the suggested unified methodology for recognizing facial expressions from images. Figure 1 illustrates the pipeline for the suggested approach schematically.

The proposed methodology is founded on the Cascaded network (DLFER), which employs a Domain Adaptation technique. DLFER comprises two deep learning models: EF-FER1 and EF-FER2, designed to tackle issues such as capturing nuanced and intricate facial cues related to emotional expressions and the lack of labeled data. The facial expression dataset (FER-2013) dataset, comprising hundreds of labeled examples, has been provided. Despite the abundance of training samples, they cannot be considered sufficient to train a deep learning model from the ground up. Consequently, this methodology has effectively utilized knowledge transfer from other domains by developing a Domain Adaptation-based DLFER inspired by EfficientNet-B3.

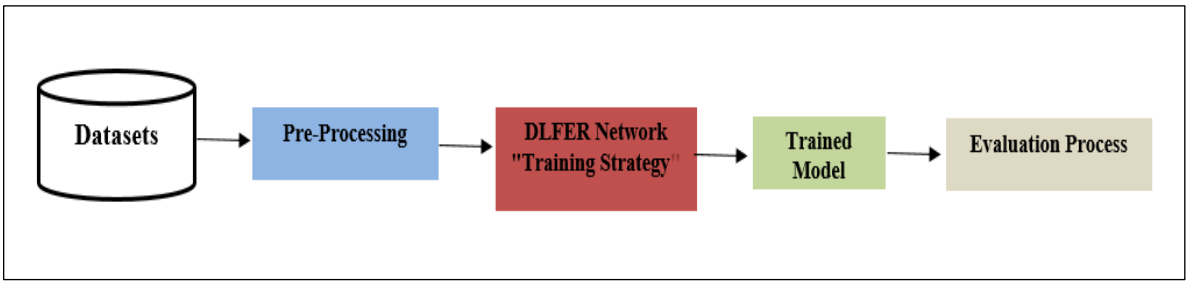

Fig. 1. The proposed approach

2.1 Data Description

In this work, two public datasets were used, namely “ImageNet” [10] and “Facial Expression Recognition 2013(FER2013)" [11]. The first public dataset for training the EF-FER1 model from scratch is an extensive collection of annotated images designed for computer vision research. It comprises 1000 classes, encompassing around one million training images.

The second public dataset is "FER2013 [Facial Expression Recognition 2013]", a publicly accessible facial expression dataset provided by Kaggle and presented at "the International Conference on Machine Learning" in 2013 by , utilized for training the EF-FER2 model. The FER2013 dataset consists of greyscale images, each with dimensions of 48 × 48. This dataset contains around 35887 greyscale photos of faces, partitioned into 28709 for training and 7178 for testing, each linked to one of seven distinct emotions: happiness, anger, surprise, fear, disgust, sorrow, and neutrality. The photographs comprise both staged and candid headshots. Figure 2 displays sample photos from the FER-2013 database.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Fig. 2. Sample images from the FER2013 dataset

The FER2013 dataset categorizes its facial images according to the seven fundamental emotions, and This variability enhances the construction and training of models utilized for emotion recognition from facial features. It is essential to take into account the inconsistency in lighting, background, and facial positioning within the image dataset attributed to variations in conditions such as camera angles and lighting settings. This complexity renders the dataset both challenging and realistic for the training and evaluation of deep learning models. However, variability in the dataset must be taken into account during the development and assessment of models.

On the other hand, a bias was observed in the dataset, favoring some facial expressions, particularly those that are more prevalent or readily identifiable. This may cause an imbalance in the distribution of facial expressions within the dataset, leading to diminished accuracy in recognizing less prevalent expressions. Table 1 presents the specifics of the FER2013 dataset, illustrating the seven fundamental emotions employed in this research.

Table 1. Distribution of emotions in training and testing the FER2013 dataset.

Emotion	Training	Testing	Total
Surprise	3171	831	4002
Fear	4097	1024	5121
Angry	3995	958	4953
Neutral	4965	1233	6198
Sad	4830	1247	6077
Disgust	436	111	547
Happy	7215	1774	8989
Total	28709	7178	35887

In addition, the FER2013 dataset poses several challenges, including a substantial quantity of low-resolution greyscale images, the presence of non-facial images, erroneous face cropping, inaccuracies in expression labeling, and ambiguity within the data, as certain images across different classes exhibit significant similarity, potentially confusing classification.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

2.2 Dataset Pre-Processing

To improve the reliability and robustness of the deep learning model, different image pre-processing techniques have been utilized in this system. The approach of image normalization, recognized as a crucial preliminary step for training deep neural networks with scientific accuracy, was adopted. This involved normalizing pixel values to a range of 0 to 255 to minimize variations.

Secondly, the dataset consists of photos with diverse width, height, and depth parameters. The photos are then resized to a uniform dimension of 224 × 224 × 1 to improve computation, as larger images necessitate increased convolutions and data processing.

Subsequently, the intensity of the pictures was normalized during the pre-processing phase. The pixel intensity of the image is normalized from its original range of 0–255 to the interval [0, 1] by uniformly dividing each value by 255, which matches the expected input range of the activation function swish, avoiding vanishing gradients and enhances the numerical stability of computations.

To enhance the model's performance and overcome the constraints of the FER2013 dataset, a balanced dataset is essential; therefore, data augmentation was executed by incorporating significantly altered replicas of the existing data. The pre-processing was conducted without modifying the essence of facial emotions. It comprised a rotation with a maximum left and right angle of ten degrees and a random zoom operation at a zoom percentage of 20%. The horizontal flip approach was employed with a probability of 0.5 to enhance symmetry while ensuring compliance with the nearest adjacent filling position. The parameters width shift range and height shift range equal to 0.1 were utilized to enrich the data by displacing photos horizontally and vertically by a maximum of 10%.

The augmentation technique yielded a balanced FER-2013 dataset; this offers a method to mitigate the bias inherent in the dataset among the current classes, hence facilitating more effective training for the deep learning model.

2.3 Proposed Cascaded DLFER Network Architecture for Facial Expression Recognition

This work presents a dataset of facial expression images with hundreds of labeled examples, referred to as the "Facial Expression Recognition 2013" Dataset. Despite the abundance of training samples, they are insufficient for training a deep learning model from scratch. So, a Cascaded Network (DLFER) for knowledge transfer from alternative domains was proposed.

DLFER is a cascaded network utilized to interpret emotions from facial expressions. The input images are analyzed to automatically extract distinguishing features by knowledge transfer from different domains, which possess abundant training data, to the Facial Expression image domain, which has a restricted training dataset for Emotion Recognition. Figure 3 illustrates the planned Cascaded Network (DLFER).

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

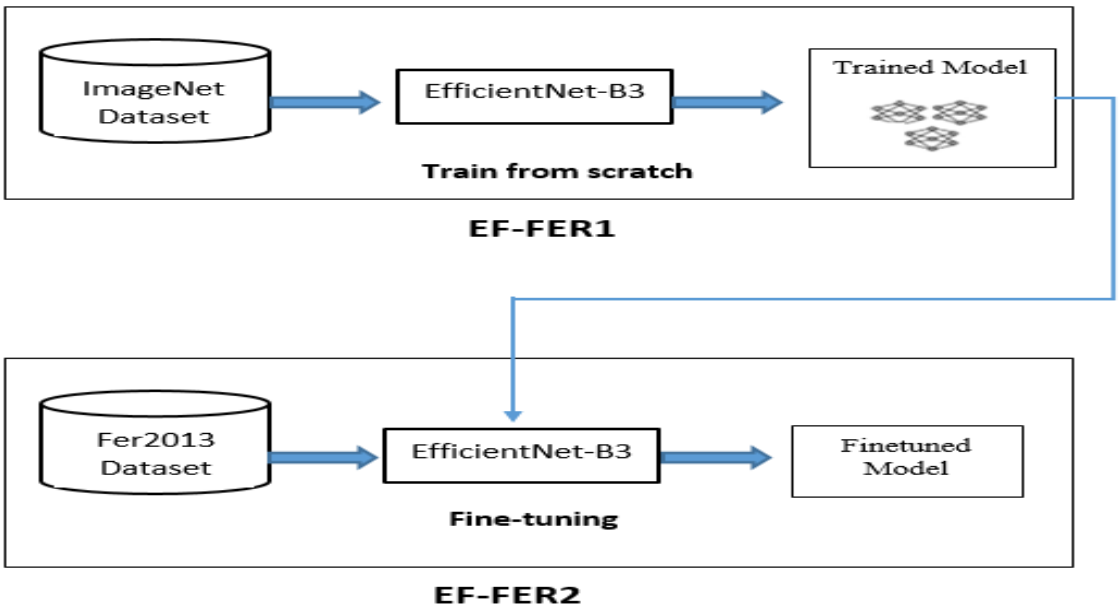

Fig. 3. Architecture of DLFER Network

The proposed DLFER network comprises two deep-learning models. The first neural model, EF-FER1, is developed and trained utilizing training data from unrelated image domains, namely ImageNet, which contains substantial training data. The learned knowledge is subsequently transferred to the secondary neural network (EF-FER2) to assist in facial expression recognition.

After training the EF-FER1 model from scratch utilizing the ImageNet dataset, the knowledge learned will be stored in the weights. With initial parameters gained from ImageNet, the DLFER network used fine-tuning to train the next model. The EF-FER2 model is initialized with weights pretrained on the "ImageNet dataset" and subsequently finetuned on the "FER2013 dataset". Consequently, the DLFER network has transferred knowledge from unrelated domains and is finetuned for emotion recognition, alleviating the limited target domain data issue.

The core of the models (EF-FER1 and EF-FER2) is EffcientNetB3 [12], famous for its exceptional performance and efficiency attributable to its compound scaling methodology. This enables the suggested method to be more accurate with reduced processing resources. The EfficientNetB3 model is trained on the ImageNet dataset to categorize 1000 image objects. In the specification of the EfficientNetB3 architecture, the last dense layers of the pretrained model are substituted with a new dense layer to classify a face image into one of seven emotional categories (happy, anger, surprise, fear, disgust, sad, and neutrality). A dense layer is a standard, fully linked linear layer that accepts an input of a specific dimension and produces an output vector of the desired dimension. Consequently, the output layer comprises solely seven neurons. The conventional fine-tuning method for the EfficientNet-B3 architecture entails updating all parameters across all layers of the pre-trained network.

EfficientNet B3 is a network model distinguished by unique characteristics. The network includes a residual structure that improves the network's depth as well as the accuracy and efficiency of feature extraction. Moreover, it allows for the alteration of the quantity of feature layers within each layer to enhance feature extraction, hence increasing the network's width. Moreover, EfficientNet-B3 can obtain and transmit information from more intricate data by augmenting the input image resolution, hence enhancing model accuracy.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

EfficientNet employs the CNN architecture by augmenting the number of layers and neurons. It consists of a stem, many blocks made up of modules, and a final layer. Table 2 illustrates the configuration of the modules.

Table 2. The structure of the EfficientNet-B3 modules

Layers	Details
stem	- Rescaling - Normalization - Zero Padding - Conv2D + Batch Normalization (BN)+ Activation
Module1	- Depth wise Conv2D + BN + Activation Function
Module2	- Depth wise Conv2D + BN + Activation function - Zero Padding - Depth wise Conv2D + BN + Activation Function
Module3	- Global average pooling - Rescaling - Conv2D + Conv2D
Module4	-Multiply - Conv2D + BN
Module5	-Multiply - Conv2D + BN+ Dropout
Final layer	- Conv2D + BN + Activation Function

The modules constitute the components of EfficientNet; they generate sub-blocks and manage the analytical process. Each layer, from module 1 to module 5, comprises five modules. Figure 4 illustrates a schematic structural depiction of the EfficientNetB3 network.. The stem layer is the initial layer that acquires the data, normalizes it, and transmits it to the modules. The final layer produces the outcome and functions as the output layer. The First Block comprises modules 1, 3, and 4. The second block consists of modules 2, 3, and 4; the final block includes modules 2, 3, and 5.Modules 3 and 4 function as the skip connections within the sub-block. Module 4 integrates the skip connection, whereas Module 5 links each sub-block to the subsequent sub-block.

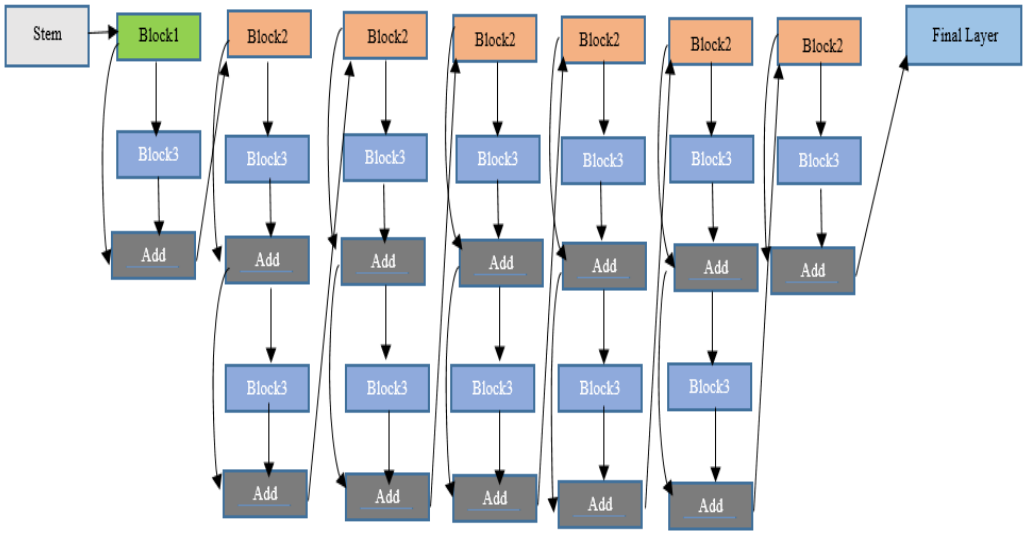

Fig. 4. The architecture of the EfficientNet-B3

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

3. EXPERIMENTATIONS

3.1 Model Training and Testing

This research used two public datasets: the "ImageNet" dataset for training the EF-FER1 model from inception and the "Facial Expression Recognition 2013 (FER2013)" dataset for training and evaluating the proposed DLFER network. The training set constituted 80% of the total images post-preprocessing, while the testing set included the remaining 20%. A validation set was derived by allocating 20% of the training set (80% training, 20% validation) due to the complexity of this model. Utilizing an extensive training set (80% of the data) when working with a complex deep neural network, such as a DLFER network with numerous layers or a model with a substantial number of parameters, can yield several advantages, including enhanced generalization.

In the training phase of this methodology, the Adam optimizer has been utilized to optimize the training weights. The primary objective is to reduce the loss function. The suggested models are trained with a learning rate of 0.001, a momentum coefficient of 0.99, a batch size of 32, and utilize binary cross-entropy loss functions. The optimization technique typically necessitates around 60 epochs to achieve convergence.

Python and the TensorFlow libraries were used to develop the codebase. The research makes use of a system with 16 GigaByte of RAM, an Intel Core i7 CPU , and a specialized graphics processing unit (GPU) with 8 GB of RAM, namely the NVIDIA GE-FORCE RTX model.

3.2 Evaluation Metrics

To assess the suggested model's efficiency, the mean of accuracy, precision, recall, and F1-score were computed, which are the conventional metrics employed by current methodologies for facial emotion identification.

Accuracy is the fundamental metric in multiclass classification. The calculation involves dividing the aggregate of true negatives and true positives for the emotion category by the total instances within that category. The mean accuracy of the classifier is calculated as:

$$\text{Accuracy}= \frac{\sum_{i=1}^{n} (\text{TPi} + \text{TNi})}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

Precision denotes the proposed model's accurate positive predictions for each emotion class. The mean precision of the model in multiclass classification is computed as

$$\text{Precision}=\frac{\sum_{i=1}^{n} \text{TPi}}{\sum_{i=1}^{n} (\text{TPi} + \text{FPi})} \quad (2)$$

Recall quantifies the proportion of positive class predictions relative to the total number of positive instances. The average recall has been calculated as follows:

$$\text{Recall}=\frac{\sum_{i=1}^{n} \text{TPi}}{\sum_{i=1}^{n} (\text{TPi} + \text{FNi})} \quad (3)$$

The F1-score is a metric that assesses the efficacy of a classifier by considering both precision and recall. An elevated F1-score signifies enhanced predictive efficacy of the categorization method. The computation can be ascertained as:

$$\text{F1-Score} = 2 \text{ x } \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

4. RESULTS AND DISCUSSION

This section presents the proposed approach's experimental study to assess the system's performance. The analysis involves comparing the proposed model's performance with prior efforts by applying their methodologies to the same public dataset. Comprehensive testing with EfficientNet-B3 shows the efficacy of the suggested emotion recognition method.

4.1 Result of Ablation Study

An ablation study provides valuable insights into the relative contribution of different architectural components to the overall performance of deep learning models. By systematically removing or altering specific components, the study evaluates how each component influences the DLFER model's effectiveness in terms of efficiency and precision.
Initially, the proposed model (EF-FER2) was trained using the Fer-2013 dataset without integrating the first EF-FER1 model into the architecture of DLFER. Subsequently, EF-FER1 was integrated into the proposed model with trained it from scratch using the ImageNet dataset.
So, the accuracy of the DLFER network can be investigated in terms of ablation result (with and without EF-FER1) in order to derive the superiority of the proposed approach. The decision to avoid training DLFER from scratch arises from the challenges associated with achieving convergence when trained from scratch, in contrast to finetuning, as illustrated in Table 3. DLFER has undergone pre-training for comparison with EF-FER1, which, conversely, is trained from scratch. In this manner, the effect of training the model has been assessed using learned features compared to training using randomly initialized weights.

Table 3. Ablation Study on Cascaded (DLFER) Network

Models	Train Accuracy	Validation Accuracy	Test Accuracy
DLFER without EF-FER1	87.80	57.74	66.21
DLFER with EF-FER1	92.31	85.26	82.09

As shown in Table 3, the proposed model improved test accuracy by 15.88% when fine-tuned with the ImageNet dataset compared to standard transfer learning, left around 18% of the samples remained misclassified. Table 4 presents a comprehensive overview of the quantitative outcomes obtained from the DLFER Model for facial emotion recognition.

Table 4. The quantitative outcomes of the DLFER Network on the test dataset

Model	**Precision**	**Recall**	**F1-score**
Proposed model (DLFER)	80.42	83.99	82.16

According to Table 4, the Model's Precision, at 80.42%, underscores its ability to reliably predict a specific emotion, reduce false positives, and produce around 19.58 % erroneous positive predictions when applied to the test set.
Furthermore, the DLFER model obtained an F1-score value of 82.16%, indicating that the model accurately identifies positive instances, minimizing the occurrence of false positives and false negatives. The accuracy and loss of the proposed model are shown in the Figures. 5(a) and 5(b) respectively.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

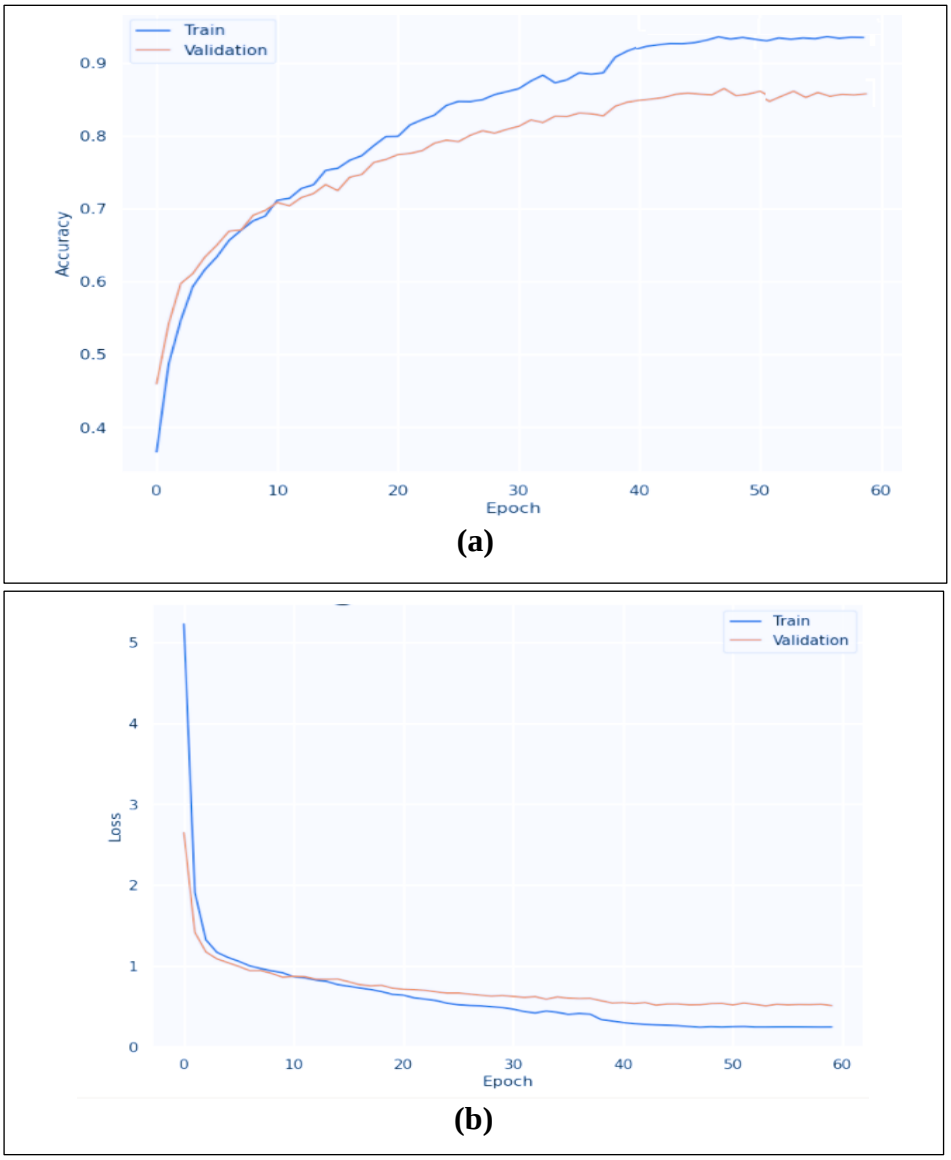

Fig. 5. (a) training versus validation accuracy and (b) training versus validation loss for the FER2013 dataset using the DLFER network

4.2 Comparison of the proposed DLFER with Prior Works:

To emphasize the importance and contribution of the work in this paper, the suggested model is compared to current achievements in the literature. The comparison results are reported in Table 5 below.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 5. Comparison of the proposed model with other deep models in previous studies On Facial expression recognition usingtheFER13dataset

Reference	Year	Architecture	Accuracy%
[13]	2020	VGGNET	72.38
[6]	2021	end-to-end deep learning	70.2
[14]	2022	CNNbasedonVGGNet	79
[15]	2022	CNN+ResNet50+Inception V3	72.3
[16]	2022	Efficient-SwishNet	63.4
[8]	2023	EfficientNet-XGBoost	72.5
[17]	2023	FER-CHC	74.68
[18]	2023	ResNet-50	73.4
[19]	2023	SSF-ViT(L)	74.95
[20]	2023	XceptionNet	77.92
[21]	2023	EmoNAS	67.9
[22]	2023	Custom CNN	62
[23]	2024	EduViTbasedontheMobileViTarchitecture	66.51
[24]	2024	shufflenetv2	65
[25]	2024	CustomCNN	57.4
[9]	2024	AlexNet + Inception V3 + ResNet50	73.56
[26]	2024	Activation-matrixTripletloss	71.62
[27]	2024	HybridizedCNN-LSTM	79.34
The proposed approach(DLFER)			**82.09**

4.3 Limitations and Future Directions :

Despite this strategy having demonstrated improved results, there are still several limitations in the current study, which influence the recognition of emotion, such as a lack of large-scale expression data, image quality, and size.

Regarding the future direction of this research, it could be expanded in several aspects: model accuracy may be further improved by taking into account existing ensemble learning methods. Additionally, future directions for this technology include exploring virtual and augmented reality applications, enhancing human-computer interaction, and integrating it with other modalities such as voice and gesture recognition.

5. CONCLUSIONS

In this study, the DLFER Network using EffcientNetB3 with a Cascaded tuning strategy has been proposed for emotion recognition from facial images. DLFER architecture, consisting of two DL models EF-FER1 and EF-FER2, working on THE FER-2013 dataset for transfer knowledge from unrelated domains, results in improved accuracy without a significant increase in computational burden. The results also showed that the DLFER model achieved better performance after being pretrained on the ImageNet dataset and achieved the highest accuracy of 82.09%.

The proposed methodology may serve as a significant instrument for its incorporation across diverse sectors, including healthcare for patient monitoring, customer service for assessing satisfaction, social sciences, and human-machine interaction.

Utilizing deep learning techniques, particularly transfer learning architectures, for emotion classification significantly enhances the progress of facial emotion identification. This study establishes a robust platform for future progress in FER. This paper addresses dataset restrictions, such as class imbalances, and explores the bounds of deep learning applications, thereby facilitating continued innovation in this exciting field.

REFERENCES

[1] 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 (Switzerland)*, vol. 10, no. 9, 2021, doi: 10.3390/electronics10091036.

[2] M. P. Sunil and S. A. Hariprasad, "Facial Emotion Recognition using a Modified Deep Convolutional Neural Network Based on the Concatenation of XCEPTION and RESNET50 V2," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 6, pp. 94–105, 2023, doi: 10.14445/23488379/IJEEE-V10I6P110.

[3] M. Kaur and M. Kumar, "Facial emotion recognition: A comprehensive review," *Expert Syst*, no. July, 2024, doi: 10.1111/exsy.13670.

[4] N. Yalçin and M. Alisawi, "Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques," *Heliyon*, vol. 10, no. 20, p. e38913, 2024, doi: 10.1016/j.heliyon.2024.e38913.

[5] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real Time Facial Expression Recognition using Deep Learning," 2019, [Online]. Available: https://ssrn.com/abstract=3421486

[6] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, pp. 1–16, 2021, doi: 10.3390/s21093046.

[7] D. Zhu, Y. Fu, X. Zhao, X. Wang, and H. Yi, "Facial Emotion Recognition Using a Novel Fusion of Convolutional Neural Network and Local Binary Pattern in Crime Investigation," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/2249417.

[8] S. B. Punuri *et al.*, "Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning," *Mathematics*, vol. 11, no. 3, pp. 1–24, 2023, doi: 10.3390/math11030776.

[9] R. K. Reghunathan, V. K. Ramankutty, A. Kallingal, and V. Vinod, "Facial Expression Recognition Using Pre-trained Architectures †," *Engineering Proceedings*, vol. 62, no. 1, pp. 4–9, 2024, doi: 10.3390/engproc2024062022.

[10] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[11] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2013, doi: 10.1016/j.neunet.2014.09.005.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[13] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11908 LNAI, pp. 273–289, 2020, doi: 10.1007/978-3-030-46133-1_17.

[14] A. Y. Nawaf and W. M. Jasim, "Human emotion identification based on features extracted using CNN," *AIP Conf Proc*, vol. 2400, no. 1, p. 20010, Oct. 2022, doi: 10.1063/5.0112131.

[15] E. G. Moung, C. C. Wooi, M. M. Sufian, C. K. On, and J. A. Dargham, "Ensemble-based face expression recognition approach for image sentiment analysis," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 2588–2600, 2022, doi: 10.11591/ijece.v12i3.pp2588-2600.

[16] T. Dar, A. Javed, S. Bourouis, H. S. Hussein, and H. Alshazly, "Efficient-SwishNet Based System for Facial Emotion Recognition," *IEEE Access*, vol. 10, no. May, pp. 71311–71328, 2022, doi: 10.1109/ACCESS.2022.3188730.

[17] X. Wu *et al.*, "FER-CHC: Facial expression recognition with cross-hierarchy contrast," *Appl Soft Comput*, vol. 145, p. 110530, Sep. 2023, doi: 10.1016/J.ASOC.2023.110530.

[18] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimed Tools Appl*, vol. 82, no. 8, pp. 11365–11394, 2023, doi: 10.1007/s11042-022-13558-9.

[19] X. Chen, X. Zheng, K. Sun, W. Liu, and Y. Zhang, "Self-supervised vision transformer-based few-shot learning for facial expression recognition," *Inf Sci (N Y)*, vol. 634, pp. 206–226, Jul. 2023, doi: 10.1016/J.INS.2023.03.105.

[20] G. Meena and K. K. Mohbey, "Sentiment analysis on images using different transfer learning models," *Procedia Comput Sci*, vol. 218, pp. 1640–1649, Jan. 2023, doi: 10.1016/J.PROCS.2023.01.142.

[21] M. Verma, M. Mandal, S. K. Reddy, Y. R. Meedimale, and S. K. Vipparthi, "Efficient neural architecture search for emotion recognition," *Expert Syst Appl*, vol. 224, p. 119957, Aug. 2023, doi: 10.1016/J.ESWA.2023.119957.

[22] L. Mozaffari, M. M. Brekke, B. Gajaruban, D. Purba, and J. Zhang, "Facial Expression Recognition Using Deep Neural Network," *2023 3rd International Conference on Applied Artificial Intelligence, ICAPAI 2023*, no. May, 2023, doi: 10.1109/ICAPAI58366.2023.10193866.

[23] L. Q. Thao *et al.*, "Monitoring and improving student attention using deep learning and wireless sensor networks," *Sens Actuators A Phys*, vol. 367, p. 115055, Mar. 2024, doi: 10.1016/J.SNA.2024.115055.

[24] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," *IEEE Access*, vol. 12, no. April, pp. 45543–45559, 2024, doi: 10.1109/ACCESS.2024.3380847.

[25] H. V. Manalu and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intelligent Systems with Applications*, vol. 21, p. 200339, Mar. 2024, doi: 10.1016/J.ISWA.2024.200339.

[26] L. Pan *et al.*, "SSER: Semi-Supervised Emotion Recognition based on Triplet Loss and pseudo label," *Knowl Based Syst*, vol. 292, p. 111595, May 2024, doi: 10.1016/J.KNOSYS.2024.111595.

[27] A. A. Bhat, S. Kavitha, S. M. Satapathy, and J. Kavipriya, "Real Time Bimodal Emotion Recognition using Hybridized Deep Learning Techniques," *Procedia Comput Sci*, vol. 235, pp. 1772–1781, Jan. 2024, doi: 10.1016/J.PROCS.2024.04.168.