Convolutional Neural Network-Based Facial Expression Recognition: Enhanced by Data Augmentation and Transfer Learning

HMLS Kumari[1#]
[1]Computer center, Faculty of Engineering, University of Peradeniya, Sri Lanka
[#]lihinisangeetha99@gmail.com

ABSTRACT Facial expression recognition has emerged as a dynamic field within computer vision and human-computer interaction, finding diverse applications such as animation, social robots, personalized banking, and more. Current studies employ transfer learning models in facial expression recognition through the application of convolutional neural networks. The proposed model combines data augmentation with fine-tunned transfer learning models to get a better FER model. A comprehensive collection of training images is crucial as input to effectively train a convolutional neural network (CNN) for accurate facial expression recognition. Hence, the presented research employed data augmentation to enhance the quantity of input images derived from a pre-existing dataset. Manually employing CNN is outdated. Therefore, fine-tuned transfer learning models are used in the proposed study. Activating the final 8 layers of the transfer learning model by freezing the whole transfer learning model is the novel methodology of the proposed model. Then we vary the values of dense layers and dropout layers of the activated 8 layers, which results the fine-tuning of the transfer learning model. The CK+, The facial recognition dataset (human) datasets are used in the proposed model. Subsequently, conduct a stratified 5-fold cross-validation to assess the model's performance on previously unseen data and avoid overfitting the proposed model. The method under consideration utilized transfer learning models, namely DenseNet121, DenseNet201, DenseNet169, and InceptionV3, along with fine-tuned transfer learning models applied to augmented datasets CK+, The facial recognition dataset (human) datasets. The outcomes indicate an achievement of 99.36% accuracy for the CK+ dataset, 95.14% for the facial recognition dataset (Human).

INDEX TERMS Accuracy, CK+, Convolutional Neural Network (CNN), Deep Learning, Data Augmentation, Facial Expression Recognition (FER), Fine-tuning, pre-trained models, Transfer learning model

I. INTRODUCTION

Facial expressions serve as a powerful and universally understood way for humans to communicate their emotions and intentions [1]. A facial expression recognition (FER) system is a computer application designed to independently identify and authenticate the emotions displayed on individuals' faces in digital images or video frames from a video feed. This is achieved by comparing the facial expressions against a database. Facial expression recognition is a significant area of contemporary research with diverse applications, including monitoring patient conditions, improving human-computer interaction, enhancing security measures, influencing game development, strengthening video surveillance capabilities, automating access control systems, animating avatars, contributing to neuro-marketing efforts, and advancing the field of sociable robots.[2]

Facial expression recognition poses challenges in the field of computer vision. because people can vary the expression of their same facial expressions in several situations [3]. As an example, people can show happy expressions differently on different occasions. Even in images of people with the same expression, the brightness, background, and pose can differ, as illustrated in Fig. 1. Therefore, facial expression recognition is a very challenging field in computer vision. As a new approach, we use a convolutional neural network with transfer learning for a small data set and then use data augmentation to make a vast dataset that is appropriate to get exceptional accuracy while fine-tuning values of the presented model.

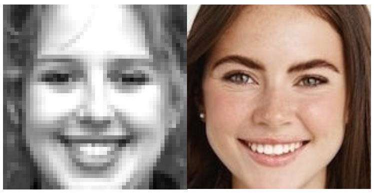


Figure 1. The two different images of a happy expression.

Fig. 1. shows that the first image is from the CK+ dataset, and the second is from The Facial Expression Dataset (Human). Even in an image of a person, the identical expression can be different in terms of brightness, background, and pose. The recognition of facial expression plays an essential role in

nonverbal communication between humans. Hence, there has been extensive research on the generation, perception, and understanding of facial expression. Therefore, the production, perception, and interpretation of facial expressions have been widely studied [4]. The universal facial expressions are happy, sad, angry, disgust, fearful, surprised, and neutral. Facial expression recognition is the main point in human emotion recognition. Darwin initiated this field of study in his book “The Expression of Emotions in Man and Animals” [1].
Recognizing expression is a task that individuals carry out effortlessly in their daily lives [5]. However, in the domain of computer vision, this is a challenging effort. There is some previous research that has different accuracy levels, such as high accuracy and low accuracy. Low accuracy is primarily caused by an uncontrolled environment, and some expressions, such as "sad" and "fear," are very similar, as illustrated in Fig (2).

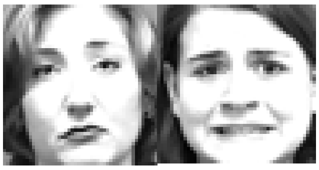


Figure 2. The "sad" and "fear" expressions are very similar

In the same dataset, "sad" and "fear" expressions are very similar, as shown in Fig. 2. As stated above, it is not an easy task in computer vision.

The use of small training datasets in research based on the classification of images leads to poor classification. A tightly constrained model may struggle to capture the details of a small training dataset, leading to underfitting. On the other hand, a loosely constrained model may excessively tailor itself to the training data, causing overfitting and ultimately resulting in inferior performance. Therefore, it is crucial to have a large dataset when training deep learning models with CNN. [7]

The deep multi-layer neural network has proven to be a successful approach in the realm of facial expression recognition. This approach integrates the three stages of facial expression recognition, such as learning, feature selection, and classification, into a single step. New research attempts to enhance the accuracy of neural networks by training them with multiple layers. But this concept results in only small increments of accuracy. While CNNs have demonstrated effectiveness in learning abstract features, especially with deeper architectures involving numerous layers and innovative training techniques [6][7][8].

Building and training a convolutional neural network manually takes time and is out of date. Therefore, one approach is using transfer learning models in convolutional neural networks. The proposed model used transfer learning models such as densenet121, densenet169, densenet201, and Inception V3.

The convolutional neural network employed in facial expression recognition demonstrates superior accuracy when applied to extensive datasets. However, one cannot easily find a dataset with a large number of images. To tackle this problem, we will use data augmentation [2][9]. In this approach, we use commonly used datasets (CK+, the facial expression (Human) dataset). The presented approach aims to attain an accuracy of 99.36% on the CK+ dataset and 95.14% on the facial expression (Human) dataset.

Facial expression recognition has improved a lot in recent decades due to the advancement of recognition methods. Deep learning, especially the improvement of convolutional neural networks, has played a key role in this progress. The effectiveness of these techniques is supported by large training datasets and ongoing improvements in GPU technology. To make small datasets more powerful, we can enlarge them through data augmentation. Because recent researchers used data augmentation in their works to increase accuracy with transfer learning models [2].

The central objective of this study is to formulate a CNN model with data augmentation and transfer learning along with CNN that achieves higher accuracy than previous works for the CK+ and the facial expression (Human) dataset which are small datasets. In the proposed study, we used per-trained models with CNN. Training the whole pre-trained model was not used in this study. Instead, activate some layers of the pre-trained model that are suitable for augmented datasets and freeze other layers. This will result in the most suitable model for each of the above datasets. We can increase accuracy by fine-tuning the values of each variable in the activated layers of the pre-trained model. Recent studies work only with data augmentation and pre-trained models with CNN and don’t use freezing layers or activate some layers of the transfer learning model.

We will demonstrate how proposed transfer learning models with fine tuning in CNN outperform recent work on the CK+ and The Facial Expression Dataset (Human) datasets with augmentation.

II. RELATED WORKS

The current facial expression recognition research shows improvement due to the rise of deep learning techniques and especially due to the evolution of convolutional neural networks. The evolution of facial expression recognition depends on reasons such as the availability of huge datasets, the ability to use and add new transfer learning methods for CNN, and the improvement of GPU technology.

Numerous recent methodologies aim to enhance accuracy in facial expression recognition. Aravind Ravi [10] investigated the utilization of features from pre-trained CNNs for facial recognition in a recent study. The findings indicate that repurposing pre-trained models designed for object recognition proves effective in facial expression recognition, with the VGG19 model's layers achieving noteworthy accuracies of 92.26% and 92.86% on the CK+ and JAFFE datasets, respectively [22]. Additionally, earlier network features exhibit high accuracy on smaller datasets, a validation achieved through 10-fold cross-validation, jack-knife validation, and leave-one-out methodologies, addressing the limitations of small datasets.

Simone Porcu, Alessandro Floris, and Luigi Atzori delved into the evaluation of data augmentation techniques for facial expression recognition systems [11]. Their study demonstrates the efficacy of data augmentation techniques in improving accuracy. Specifically, geometric data augmentation and generative adversarial networks contribute to a 30% increase in CNN accuracy using the VGG16 architecture. Employing these methods successfully expands the training dataset initially based on the KDEF dataset and subsequently tests its efficacy on the CK+ and Expw datasets.

Narayana Darapaneni, Rahul Choubey, and Pratik Salvi conducted an investigation into facial expression recognition and recommendations using deep neural networks with transfer learning [12]. The study employed the Jaffe dataset and utilized VGG-16 and InceptionV3 as two transfer learning models[22]. Training configurations, including the last 5 layers, the last 3 layers, the last 1 layer, and all layers, were explored with recognition rates of 95% and 94% achieved through cross-validation.

In another exploration, Tawsin Uddin Ahmed, Sazzad Hossain, Mohammad Shahadat Hossain, Raihan Ul Islam, and Karl Andersson delved into facial expression recognition using a convolutional neural network with data augmentation [13]. This study showcased the effectiveness of data augmentation in enhancing CNN accuracy. Datasets such as CK+, FER 2013, the MUG facial expression database, KDEF and AKDEF, and KinFaceW-I and II were employed, resulting in an overall CNN accuracy of 95.87%.

Andre Teixeira Lopesa, Edilson de Aguiarb, Alberto F. De Souzaa, and Thiago Oliveira-Santosa explored facial expression recognition with convolutional neural networks, specifically addressing the challenges of limited data and training sample order [14]. Small datasets, including CK+, JAFFE, and BU-3DFE, were augmented to create substantial datasets for deep architecture-based facial expression recognition. The proposed method achieved an impressive 96.76% accuracy on the CK+ dataset [22].

The facial expression recognition model has achieved high accuracy by forming an ensemble of modern deep CNNs. Christopher Pramerdorfer and Martin Kampel conducted research and obtained 75.2% accuracy for the FER2013 dataset [15]. They used CNN architectures such as VGG16, Inception, and Resnet. They perform a thorough search to identify the best ensembles of up to 8 models in terms of FER2013 validation accuracy. Real-time facial expression recognition using deep learning research was proposed by Isha Talegaonkar and team [16]. They used the FER2013 dataset and made different changes for the number of epochs, number of layers, and layers of the CNN architecture to produce the model with the highest accuracy. As a result, they achieve a training accuracy of 79.89% and a test accuracy of 60.12% for the FER2013 dataset.

III. METHODOLOGY

A. Dataset

The datasets used in this study are CK+ and the facial recognition dataset (human). The CK+ dataset is the Extended Cohn-Kanade dataset, which contains 123 different subjects and their 593 video sequences [17]. The images in the dataset are of people whose ages range from 18 to 50 and represent a variety of genders and heritages. The CK+ contains 327 labeled images with seven facial expression classes: anger, disgust, contempt, fear, happiness, sadness, and surprise. All images in the CK+ dataset are grayscale images. This dataset is widely used in facial expression classification. The number of images in each class of the dataset CK+ is shown in Table 1. The number of images in classes is different in the CK+ dataset, as shown in Table 1. Therefore, we have considered the unbalance of the ck+ dataset when building the proposed model with the CNN model. Table 1 shows the classes of the ck+ dataset and their number of images.

Table 1. Emotion and number of images in each class of ck+ dataset is represented

Emotion	**Number of images**
Angry (An)	45
Contempt (Co)	18
Disgust (Di)	59
Fear (Fe)	25
Happy (Ha)	69
Sadness (Sa)	28
Surprise (Su)	83

The facial recognition dataset (human) consists of 1823 images and represents 5 facial expressions [18]. The classes of the dataset are: angry_human_face, happy_human_face, neutral_human_face, sad_human_face, and surprised_human_face. The imbalance of classes cannot be seen in this dataset. The images are not grayscale. The whole dataset was divided into ratios of 80%, 10%, and 10% for training, testing, and validation, respectively. Table 2. shows the number of images in each class of facial expression (human) dataset.

The sample images that show different classes of datasets (CK+ and Facial Expression (human)) are shown in Figs. 3 and 4 below, respectively.

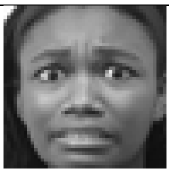
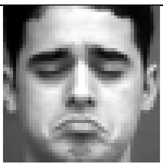
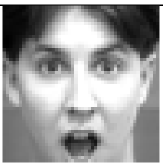
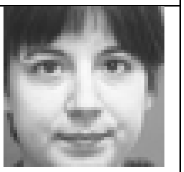
	Angry	Disgust	Fear	Happy	Sad	Suprise	Neutral
CK+							

Figure 3. The sample images that show different classes of CK+ dataset.

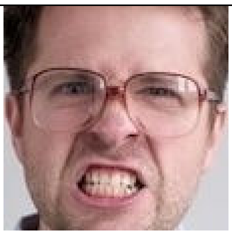
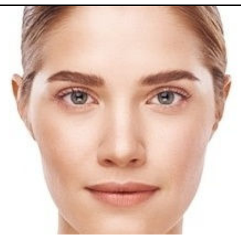
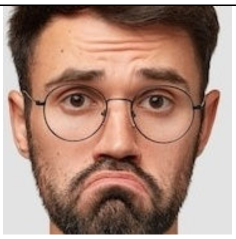
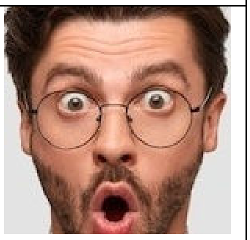
	Angry human face	Happy human face	Neutral human face	Sad human face	Surprised human face
Facial Expression (Human) dataset					

Figure 4. The sample images that show different classes of Facial expression (Human) dataset.

Table 2. Classes of each dataset and their number of images in The Facial Expression (Human) dataset

Expression class	**Facial Expression (Human) dataset**
Angry_human_face	355
Happy_human_face	410
Neutral_human_face	367
Sad_human_face	308
Surprised_human_face	383

B. CNN and Transfer Learning models

A convolutional neural network (CNN) is a specialized type of artificial neural network designed mainly for image recognition within the broader realm of deep learning. CNNs are particularly effective at analyzing pixel data and identifying intricate patterns in images [19]. The process involves taking an image as input, recognizing important learnable weights and biases related to different objects in the image, and allowing the network to distinguish between distinct objects. The CNN architecture consists of four key layers: the convolutional layer, pooling layer, RELU-connection layer, and fully connected layer. This combination makes CNNs well-suited for tasks like facial expression recognition, making them prominent in recent studies exploring the complexities of facial expressions.

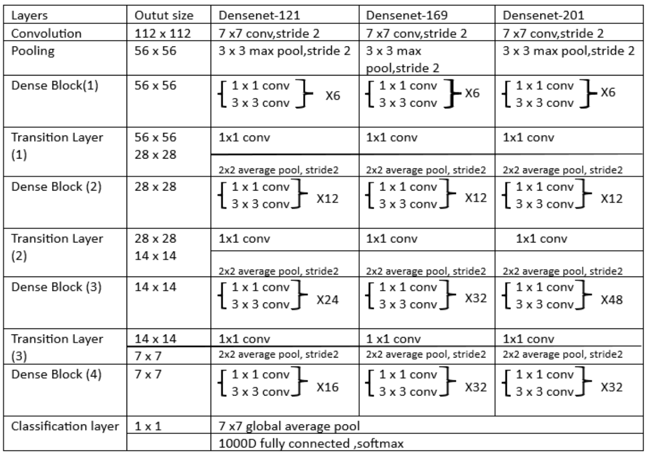
A popular way to recognize facial expressions using CNNs is by using transfer learning models with pre-trained weights from Keras applications. These advanced models are used for tasks like prediction, fine-tuning, and feature extraction in facial expression recognition through CNNs. [26] Many Keras models, such as Densenet121, Densenet169, Densenet201, and Inceptionv3, have been recently used in studies for this purpose.

Fine-tuning stands out as a popular transfer learning technique, particularly for achieving effective facial expression recognition on diverse datasets using pre-trained CNNs. [15] The fine-tuning process involves three key steps:

- Adapt the pre-trained network by eliminating its final layer (the softmax layer) and substituting it with a new softmax layer customized for our particular model.Since pre-trained networks are designed for a larger number of categories, typically 1000 or more, adaptation is necessary for our task of classifying seven facial expressions. Cross-validation is employed to ensure the proper functioning of the adapted network.
- During the training of the pre-trained CNN model with the dataset, a small learning rate is utilized to enhance the model's adaptability.
- Certain layers of the pre-trained network are frozen, and new layers are introduced to align with the characteristics of our dataset.

This study incorporates all these fine-tuning methods, employing various transfer learning models such as Densenet121, Densenet120, Densenet169, and Inception V3.

Densenet is a deep learning network renowned for its efficiency in training, employing concise connections linking every layer.



Layers	Outut size	Densenet-121	Densenet-169	Densenet-201
Convolution	112 x 112	7 x7 conv,stride 2	7 x7 conv,stride 2	7 x7 conv,stride 2
Pooling	56 x 56	3 x 3 max pool,stride 2	3 x 3 max pool,stride 2	3 x 3 max pool,stride 2
Dense Block(1)	56 x 56	{1 x 1 conv 3 x 3 conv} X6	{1 x 1 conv 3 x 3 conv} X6	{1 x 1 conv 3 x 3 conv} X6
Transition Layer (1)	56 x 56 28 x 28	1x1 conv 2x2 average pool, stride2	1x1 conv 2x2 average pool, stride2	1x1 conv 2x2 average pool, stride2
Dense Block (2)	28 x 28	{1 x 1 conv 3 x 3 conv} X12	{1 x 1 conv 3 x 3 conv} X12	{1 x 1 conv 3 x 3 conv} X12
Transition Layer (2)	28 x 28 14 x 14	1x1 conv 2x2 average pool, stride2	1x1 conv 2x2 average pool, stride2	1x1 conv 2x2 average pool, stride2
Dense Block (3)	14 x 14	{1 x 1 conv 3 x 3 conv} X24	{1 x 1 conv 3 x 3 conv} X32	{1 x 1 conv 3 x 3 conv} X48
Transition Layer (3)	14 x 14 7 x 7	1x1 conv 2x2 average pool, stride2	1 x1 conv 2x2 average pool, stride2	1x1 conv 2x2 average pool, stride2
Dense Block (4)	7 x 7	{1 x 1 conv 3 x 3 conv} X16	{1 x 1 conv 3 x 3 conv} X32	{1 x 1 conv 3 x 3 conv} X32
Classification layer	1 x 1	7 x7 global average pool		
		1000D fully connected ,softmax		

Figure 5. Architecture of Densenet121, Densenet169, and Densenet201

In the diagram Fig.5, shows that every Densenet model comprises four dense layer blocks, each with varying numbers of layers. This discrepancy in the number of layers is the key distinction among densenet121, densenet169, and densenet201. [25] [26]

InceptionV3 architecture:

The inception architecture consists of several concepts. Factorized convolution can be seen. This checks network efficiency. The second concept is small convolutions. It replaced a large convolution with small convolutions. It leads to faster training. Next: asymmetric convolutions. It replaces 3x3 convolutions by 1x3 convolutions, followed by 3x1 convolutions. An auxiliary classifier is a small layer insert between layers. This is a small CNN layer. The final concept is grid size reduction, which is done by pooling operations. This will make a model more efficient and avoid computational costs. All these concepts combine into one model and form Inception V3. [27]

C. Data Augmentation

Addressing computer vision tasks, such as facial expression recognition, with a limited training set poses a significant challenge for CNNs. As a result, we must investigate whether there is an increase in accuracy with dataset augmentation when using transfer learning modelsData augmentation is used to make an extensive training dataset suitable for facial expression recognition using CNN. Data augmentation methods were crop, flip, Gaussian blur, contrast normalization, additive Gaussian noise, scale, multiply, translate percent, shear, and rotate. The sample images of the CK+ dataset and the facial expression (human) dataset with the data augmentation showed in Fig. 6, Fig.7.

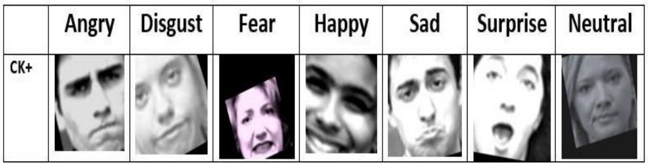


Figure 6. The sample images of CK+ dataset with the data augmentation.

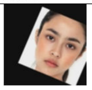
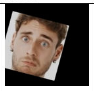
	Angry human face	Happy human face	Neutral human face	Sad human face	Surprised human face
Facial Expression (Human) dataset					

Figure 7. The sample images of Facial expression (Human) dataset with the data augmentation.

D.The Proposed FER System

First, split the dataset as shown in the above diagram. Divide the whole dataset into three parts: the training dataset (70%), the test dataset (20%), and the validation dataset (10%) from the whole dataset. Take the train dataset and apply data augmentation. The data augmentation is done by synthesizing one image into 10 images in the training dataset. In this study, some data-augmentation methods are used. Those are flipping, Gaussian blur, linear contrast, multiplying the number of images, scaling, translating percent, and rotating. This step was done to increase the training dataset and avoid poor classification. Because the model extracts all features and other necessary information using a training dataset to classify test data correctly, Then send those images and their labels separately to the list. Then we have to make a CNN model using a transfer learning model. Transfer learning models are pre-trained for classification tasks using an extensive number of images. Therefore, it is easy to change those transfer learning models to classify similar tasks, such as facial expressions, that are present in the test dataset. The first step of the procedure is to import the transfer learning model.

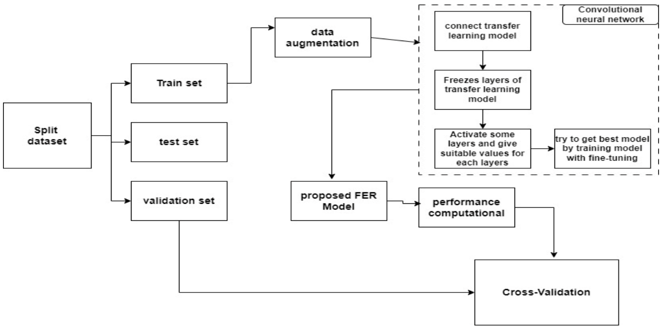

Split dataset
Train set
test set
validation set
data augmentation
Convolutional neural network
connect transfer learning model
Freezes layers of transfer learning model
Activate some layers and give suitable values for each layers
try to get best model by training model with fine-tuning
proposed FER Model
performance computational
Cross-Validation

Figure 8. The diagram shows proposed FER system

The subsequent step involves freezing all layers of the transfer learning model to reduce the risk of overfitting and prevent training the entire network. Unfreezing the final eight layers is then performed to capture detailed information in the images, such as image edges. Following these adjustments, the model demonstrates a good fit and can further be fine-tuned by adjusting the variable values in the layers.

The subsequent stage involves putting the test images, validation images, and label list into the CNN model and executing cross-validation on the test dataset. Cross-validation is used to estimate the new model's behavior for new data (images and data). In this study, it used five stratified cross-validations. The five-stratified cross validation maintains proportions of classes in each fold and prevents overfitting of the new proposed model.

The proposed FER model shows in Fig.(8) results in a new model with high accuracy to classify facial expression, and importantly, the model originates from small datasets.

E. Training

Training involved the utilization of two datasets separately, the CK+ and facial expression (human) datasets, with the incorporation of data augmentation techniques. In this investigation, we introduced eight additional layers by maintaining the immobility of all transfer learning model layers [15]. These augmentations encompassed a GlobalAveragePooling2D layer, two dropout layers, two dense layers, and one batch normalization layer. Hyperparameter tuning was conducted by varying the dropout layer values within the range of 0.4 to 0.7 and experimenting with dense layer configurations, specifically 1024, 512, and 128. The training process spanned 30 epochs, employing diverse batch sizes of 16, 32, and 64. Ultimately, a 5-fold stratified cross-validation methodology was employed across the CK+ and facial expression (human) datasets, integrating multiple transfer learning models to identify the most optimal models.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Implementation Details

In the proposed method, we used densenet121, densenet210, densenet169, and inceptionV3 for images in CK+ and the facial expression (human) dataset. The image size has been set to 224 x 224. In this proposed model, EarlyStopping, ModelCheckPoint, and ReduceLROnPleateau were used. [8]. The model monitors the accuracy.

The proposed model trains for 30 epochs and for batch sizes 16, 32, and 64. We used Google Colab with Python Language and Keras Libraries that run on Tenserflow Basement in this study. The Google colab environment has access to the NVIDIA Tesla K80.

B. Evaluation metrics

Accuracy, Precision and F1 score are the evaluation metrics of this study.

$$\text{Accuracy} = {(TP + TN)}\big/{(TP + TF + FP + FN)}$$

$$\text{Precision} = {TP}\big/{(TP + FP)}$$

$$\text{Sensitivity} = {TP}\big/{(TP + FN)}$$

TP=true positive, TN=true negative, FP=false positive, FN=false negative

Accuracy shows how often a facial expression classification model is correct overall. Precision shows how often a proposed facial expression model is correct when predicting the target class. Recall shows whether the proposed facial expression model can find all images of targeted facial expressions. [23][24]

C. Experimental setup

Cross-validation is used in this proposed FER model. We need to measure how the proposed model behaves in the presence of unseen images. Stratified 5-fold cross-validation was used in this study. This cross-validation type is an extension technique used for classification problems. Mainly, this is because the CK+ and the facial expression (human) datasets are imbalanced. Therefore, we need to keep the same proportion of classes throughout the k-folds as the original dataset.

D. Testing Results

The present investigation assesses the performance using the facial expression (human) and CK+ datasets, which are commonly employed in facial emotion recognition (FER) research due to their compact size. A comparative analysis is conducted with recent studies, demonstrating the contemporary nature of our approach and its commendable accuracy on datasets like CK+ and the facial expression (human) datasets. Our methodology leverages popular transfer learning models, including Densenet121, Densenet201, Densenet169, and Inception V3, which currently dominate the landscape of FER systems. Notably, existing studies often neglect the combined application of data augmentation, transfer learning, and fine-tuning for achieving optimal accuracy. The subsequent results provide a detailed breakdown of the accuracy achieved by our proposed model, separately employing Densenet121, Densenet169, Densenet201, and Inception V3 on the CK+ and the facial expression (human) datasets.

Table 3. Final Maximum accuracies gained by proposed model for ck+ dataset

Model	Batch size	Dense layers	Drop out value	Accuracy	Precision	F1 Score
Densenet 121	32	1024, 128	0.4	0.9937	0.9937	0.9937
Densenet 169	32	1024, 128	0.4	0.9905	0.9910	0.9906

Densenet 201	32	1024, 128	0.5	0.9842	0.9855	0.9843
InceptioV 3	32	1024, 128	0.4	0.9645	0.9675	0.9336

TABLE 4. Final Maximum accuracies gained by proposed model for Facial Expression (Human) dataset

Model	Batch size	Dense Layers	Drop out Value	Accuracy	Precision	F1 Score
Densenet121	32	1024, 512	0.5	0.9274	0.9283	0.9277
Densenet201	32	1024, 512	0.5	0.9514	0.9517	0.9514
Densenet169	32	1024, 128	0.4	0.9139	0.9169	0.9142
InceptioV3	32	1024, 128	0.4	0.8879	0.8801	0.8840

According to the tables 3,4 , the proposed model attained a peak accuracy of 99.37% for CK+ and 95.14% for the facial expression (human) dataset. The best accuracies of the above models are shown using the graph shows in Fig.9.

Table 5 shows how the proposed model achieves the best results with respect to previous work. Most effective accuracy can be achieved by combining data augmentation and transfer learning models with new layers along CNN.

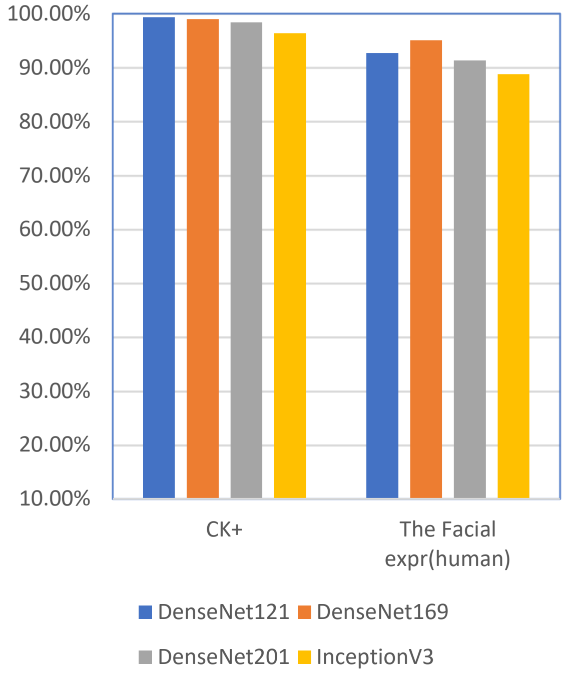


Figure 9. The graph shows maximum accuracy with each model

TABLE 5. comparison of proposed method and previous models' accuracy for CK+, the facial expression (human) datasets.

V. CONCLUSION

Study	Dataset	Accuracy
Aravind Ravi performs a study regarding Pre-Trained CNN Features for FER [10][22]	CK+, JAFFE	92.26% 92.86%
Evaluation of Data Augmentation Techniques for FER Systems by Simone Porcu, Alessandro Floris and Luigi Atzori [11]	CK+ ExpW	83.30%
Narayana Darapaneni, Rahul Choubey, Pratik Salvi did a study on FER and Recommendations Using Deep Neural Network with Transfer Learning [12][22]	JAFFE VGG16 InceptionV3	95% 94%
Facial Expression Recognition using Convolutional Neural Network with Data Augmentation by Tawsin Uddin et al.[13]	CK+, FER 2013, The MUG Facial expression database,etc	95.87%
Facial Expression Recognition with CNN: with coping with few data and the training sample order study done by Andre Teixeira Lopesa et al.[14][22]	CK+, JAFFE BU-3DFE	96.76%
Facial Expression Recognition using CNN: State of the Art by Christopher Pramerdorfer et al.	FER2013	75.2%
Proposed FER model	**CK+, Facial Expression (Human) dataset**	**99.37%** **95.14%**

In this research, a contemporary approach to facial expression recognition was introduced, employing a CNN architecture coupled with a transfer learning model and data augmentation. Noteworthy pre-trained models, including DenseNet121, DenseNet201, DenseNet169, and InceptionV3, commonly utilized in image classification, were incorporated. The study demonstrates the enhanced efficiency of classification achieved through the fine-tuning of transfer learning models.

Despite the limitations of small datasets such as CK+ and the facial expression (human) dataset, known for their modest size and limited responsiveness, our methodology leveraged data augmentation to augment the dataset size.

The core idea of transfer learning is simple. Utilize a model trained on a large dataset and apply its knowledge to a smaller dataset. In facial expression recognition with a CNN, we freeze the initial convolutional layers and only fine-tune the last 8 layers responsible for prediction.

The reasoning behind this approach lies in the fact that convolutional layers capture general, fundamental features applicable across diverse images, like edges, patterns, and gradients. Subsequent layers then specialize in recognizing specific features within an image, such as eyes or noses. By applying transfer learning models in conjunction with fine-tuning, the study successfully addressed the challenges posed by small datasets. As evident in the aforementioned results, this approach emerges as the optimal solution for facial expression recognition systems employing convolutional neural networks on small datasets, showcasing the synergistic impact of data augmentation, transfer learning, and fine-tuning.

REFERENCES

[1] "University of Glasgow," The expression oftheemotionsinmanandanimals,https://www.gla.ac.uk/myglasgow/library/files/special/exhibns/month/nov2009.html (accessed Nov. 24, 2023).

[2] Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215. https://doi.org/10.1109/taffc.2020.2981446

[3] Fathallah, A., Abdi, L., & Douik, A. (2017). Facial expression recognition via deep learning. 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). https://doi.org/10.1109/aiccsa.2017.124

[4] Jia, S., Wang, S., Hu, C., Webster, P. J., & Li, X. (2021). Detection of genuine and posed facial expressions of emotion: Databases and methods. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.580287

[5] Handbook of Face Recognition.(2011) https://doi.org/10.1007/978-0-85729-932-1

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

[7] T. U. Ahmed, S. Hossain, M. S. Hossain, R. ul Islam, and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation," 2019 Joint 8th International Conference on Informatics, Electronics &amp;amp; Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision &amp;amp; Pattern Recognition (icIVPR), 2019. doi:10.1109/iciev.2019.8858529

[8] S. Alizadeh and A. Fazel, "Convolutional neural networks for facial expression recognition," [1704.06756] Convolutional Neural Networks for Facial Expression Recognition, http://export.arxiv.org/abs/1704.06756 (accessed Nov. 24, 2023).

[9] S. Porcu, A. Floris, and L. Atzori, "Evaluation of data augmentation techniques for facial expression recognition systems," *Electronics*, vol. 9, no. 11, p. 1892, 2020. doi:10.3390/electronics9111892

[10] A. Ravi, "Pre-trained convolutional neural network features for facial expression recognition," arXiv.org, https://arxiv.org/abs/1812.06387 (accessed Nov. 27, 2023).

[11] S. Porcu, A. Floris, and L. Atzori, "Evaluation of data augmentation techniques for facial expression recognition systems," Electronics, vol. 9, no. 11, p. 1892, 2020. doi:10.3390/electronics9111892.

[12] N. Darapaneni *et al.*, "Facial expression recognition and recommendations using deep neural network with transfer learning," *2020 11th IEEE Annual Ubiquitous Computing, Electronics &amp; Mobile Communication Conference (UEMCON)*, 2020. doi:10.1109/uemcon51285.2020.9298082

[13] Md. Z. Uddin, W. Khaksar, and J. Torresen, "Facial expression recognition using salient features and convolutional neural network," *IEEE Access*, vol. 5, pp. 26146–26161, 2017. doi:10.1109/access.2017.2777003

[14] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017. doi:10.1016/j.patcog.2016.07.026

[15] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," arXiv.org, https://arxiv.org/abs/1612.02903v1 (accessed Nov. 24, 2023).

[16] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok, and A. Kulkarni, "Real time facial expression recognition using deep learning," *SSRN Electronic Journal*, 2019. doi:10.2139/ssrn.3421486

[17] *Papers with code - CK+ dataset*. CK+ Dataset | Papers With Code. (n.d.). https://paperswithcode.com/dataset/ck

[18] Khan, Z. (2023, November 24). *Facial recognition dataset (human)*. Kaggle. https://www.kaggle.com/datasets/zawarkhan69/human-facial-expression-dataset

[19] Awati, R. (2023, April 24). *What are convolutional neural networks?: Definition from TechTarget*. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network

[20] X. Wang, K. Wang, and S. Lian, "A survey on Face data augmentation for the training of Deep Neural Networks," *Neural Computing and Applications*, vol. 32, no. 19, pp. 15503–15531, 2020. doi:10.1007/s00521-020-04748-3

[21] Stanford University CS231N: Deep Learning for Computer Vision, http://cs231n.stanford.edu/reports/2016/pdfs/023_Report.pdf (accessed Nov. 24, 2023).

[22] Lyons, Michael, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. "Coding facial expressions with gabor wavelets." In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pp. 200-205. IEEE, 1998.

[23] Accuracy vs. precision vs. recall in machine learning: What's the difference? Evidently AI - Open-Source ML Monitoring and Observability. (n.d.). https://www.evidentlyai.com/classification-metrics/accuracyprecisionrecall#:~:text=Accuracy%20shows%20how%20often%20a,objects%20of%20the%20target%20class.

[24] Mage.ai. (n.d.). https://www.mage.ai/blog/definitive-guide-to-accuracy-precision-recall-for-product-developers

[25] A. Ahmed, "Architecture of densenet-121," OpenGenus IQ: Computing Expertise &amp; Legacy, https://iq.opengenus.org/architecture-of-densenet121/ (accessed Nov. 24, 2023).

[26] G. Singhal, "Gaurav Singhal," Pluralsight, https://www.pluralsight.com/guides/introduction-to-densenet-with-tensorflow (accessed Nov. 20, 2023).

[27] V. Kurama, "A guide to resnet, inception V3, and squeezenet," Paperspace Blog, https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet/ (accessed Nov. 24, 2023)

[28] Ramalingam, S., & Garzia, F. (2018). Facial expression recognition using transfer learning. 2018 International Carnahan Conference on Security Technology (ICCST). https://doi.org/10.1109/ccst.2018.8585504

[29] Randellini, E., Rigutini, L., & Saccà, C. (2021). Data Augmentation and transfer learning approaches applied to facial expressions recognition. *NLP Techniques and Applications*. https://doi.org/10.5121/csit.2021.111912

[30] Darapaneni, N., Choubey, R., Salvi, P., Pathak, A., Suryavanshi, S., & Paduri, A. R. (2020). Facial expression recognition and recommendations using deep neural network with transfer learning. *2020 11th IEEE Annual Ubiquitous Computing, Electronics &amp; Mobile Communication Conference (UEMCON)*. https://doi.org/10.1109/uemcon51285.2020.9298082

[31] Ahmed, T. U., Hossain, S., Hossain, M. S., ul Islam, R., & Andersson, K. (2019). Facial expression recognition using convolutional neural network with data augmentation. *2019 Joint 8th International Conference on Informatics, Electronics &amp; Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision &amp; Pattern Recognition (icIVPR)*. https://doi.org/10.1109/iciev.2019.8858529

[32] Hrga, I., & Ivasic-Kos, M. (2022). Effect of data augmentation methods on face image classification results. *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. https://doi.org/10.5220/0010883800003122