A Comparative Study: Facial Emotion Recognition by Using Deep Learning

1st Gan Yip Fong
Faculty of Information Science & Technology
Multimedia University
Melaka, Malaysia
joskenfong234@gmail.com

2nd Dr Goh Pey Yun
Faculty of Information Science & Technology
Multimedia University
Melaka, Malaysia
pygoh@mmu.edu.my

3rd Dr Choong Lee Ying
Faculty of Information Science & Technology
Multimedia University
Melaka, Malaysia
lychong@mmu.edu.my

4th Dr Tan Shing Chiang
Faculty of Information Science & Technology
Multimedia University
Melaka, Malaysia
sctan@mmu.edu.my

2024 12th International Conference on Information and Communication Technology (ICoICT) | 979-8-3503-6343-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOICT61617.2024.10698120

***Abstract*—Facial emotion recognition (FER) has garnered significant attention due to its applications in human-computer interaction, healthcare, psychology, and entertainment. Deep learning algorithms have been widely applied and improve the performance of FER. However, some critical gaps in FER, including the lack of standardized comparative platforms, limited used of evaluation metrics may have hinder the comprehensive understanding of various deep learning algorithms in FER. This study proposed some useful guidelines which include a standardized platform, more evaluation metrics and explored the impact of transfer learning and input size on FER. These datasets: CK, FER aligned, FER aligned + CK, RAF-DB, and AffectNet are used in the experiment. The research used the cloud-based platform, i.e. Kaggle as the standardized platform and the code was released through GitHub. Through the suggested guidelines for evaluation, some interesting insights such as the performance of each different emotion for each dataset and each deep learning algorithm, the impact of different input sizes and transfer learning are revealed.**

Keywords—facial emotion recognition (FER), deep learning algorithms, transfer learning, input dimensions

I. INTRODUCTION

Facial emotion recognition (FER) was the process of recognizing and interpreting a person's emotions from facial expression, attracting much attention in areas including human-computer interaction [1], healthcare medicine psychology [2], and entertainment [3]. This complex process involves studying a variety of facial features like muscle movements and expressions to determine the emotions. Several major facial emotions are universally recognized across cultures, including anger, disgust, fear, happiness, sadness, surprise, and neutral [4]. These emotions, according to [5], are essential components of human connection and are crucial for social communication and comprehension.

Technology advancement in computer vision and deep learning algorithms have largely accelerated the improvement of emotion recognition and analysis in terms of classification accuracy. The input data may be images or videos. However, there are several challenges that have been found in the field of the FER. Although many FER algorithms have been proposed, there is a scarcity in terms of platforms for facial emotion where FER related algorithms could hardly be evaluated and benchmarked to understand the ability and efficiency of each algorithm [6], [7]. This may hinder the improvement progress in FER algorithm.

Another significant challenge is the inadequate use of evaluation metrics [8]. Most studies focus primarily on classification accuracy, but this singular metric may not provide a comprehensive assessment of an algorithm's performance. Metrics such as recall is crucial for understanding different aspects of an algorithm's effectiveness, especially in scenarios with imbalanced datasets of class. In other words, there is a lack of comprehensive understanding of the effectiveness of various deep learning algorithms in FER.

In addition, based on [9], increasing the image input size may contribute to better comprehension of the machine learning performance in FER. These challenges motivate this study to examine the gaps in the field of FER concerning the lack of publicly available code, lack of standardized platforms for comparative analysis, the limited use of evaluation metrics and the impact of input dimensions on the performance of deep learning algorithms. There are also researchers claimed that transfer learning [10]-[12] are impacting the performance of deep learning algorithms. Addressing these obstacles can provide more insights to the other interested researchers as a guideline in using the standard platform, evaluating and assessing a new FER algorithm.

In this study, we attempted to fill up the gap by suggesting the following guideline as the initial assessment and evaluate few deep learning algorithms so that a more convincible way is shaped to help researchers and developers make informed choices about the performance of the deep learning algorithm in FER:

1. A cloud-based platform that can be easily used by all the other researchers as a standard platform and the release of code for benchmarking purposes.
2. More evaluation metrics besides classification accuracy where this includes the recall. These metrics help understand model performance across different classes, especially in imbalanced datasets, leading to a more accurate evaluation.
3. The impact of transfer learning and input dimensions are explored in this comparative study.

By demonstrating these guidelines through experimental analysis, this study provides a more reliable framework to help researchers and developers make informed choices about the performance of deep learning algorithms in FER.

II. Related Works

Several methods have been developed for FER using deep learning that relies on the architecture modification of convolutional neural networks (CNNs). [13] showed that the use of CNN models such as VGG16 and ResNet50, along with global average pooling for feature extraction and a multi-layer perceptron (MLP) for classification, can achieved high accuracy in emotion detection in facial images. [14] developed an attention CNN whose architecture comprises convolutional layers, spatial transformer modules, and visualization techniques that helped to determine the critical facial regions for emotion recognition. Good performance was achieved on FERG, JAFFE, and FER datasets.

There are researchers who applied pre-trained technique, i.e. transfer learning to train the deep learning model. For example, [10] pre-trained and fine-tuned AlexNet for better performance, yielding high accuracy on CK+ and FER datasets. [11] and [12] proposed fine-tuning of pre-trained CNN architectures, i.e. VGG and ResNet50 for efficient facial recognition across challenging datasets, such as RAF-DB and FER-2013.

Some works explored the impact of pre-processing techniques in FER. [15] highlighted the role of pre-processing techniques such as cropping, resizing and intensity normalization for enhancing the accuracy of the outcome with the CNN models. [16] integrated pre-processing techniques such as cropping, rescaling, and augmentation into the Ad-Corre model for FER-2013, RAF-DB, and AffectNet datasets. [17] proposed a Siamese Neural Network model that used DenseNet121 for extracting the image features and applied pre-processing and optimization of the loss functions, which proved to be an efficient way of utilizing multiple signals for emotion recognition. These research works showed that CNN-based models with pre-processing techniques are efficient in classifying facial emotion samples.

In general, these research works have shown the effectiveness of deep learning architecture but with different platform and comparison manner. ResNet50 was included in this comparative study due to the high accuracy in FER [10]. Other recent famous deep learning models include DenseNet (achieve very good performance but with fewer parameters), InceptionV3 or known as GoogleNet (has the strengths in solving deeper network with wider model through multi-scale processing), and Xception (the improved version of InceptionV3 which further enhance the efficiency of the algorithm) [18] were applied in this comparative study to understand the effectiveness of the deep learning algorithms in FER. In addition, the related works [15] – [17] inspired the inclusion of pre-processing techniques (including as data augmentation, grayscale conversion and one-hot encoding) and transfer learning to explore the impact on deep learning algorithms in FER. The famous datasets include CK, FER, RAFDB, and AffectNet were used in this study.

III. Methodology

In this section, the datasets, pre-processing transfer learning, model architecture, measurement metrics, and process flows of the models are explained.

A. Datasets

Table I shows the details of datasets in this study. The number of each emotion sample was reported under 'Emotions. The dimensions or image size and number of channels (i.e. 3 is for RGB and 1 is for grayscale) were reported under 'Input Shape'.

TABLE I. Datasets Used

Dataset	Emotions	Total Images	Input Shape	Purpose
CK Dataset	Surprise (249), Fear (75), Angry (135), Sadness (84), Happy (207)	750	(48, 48, 1)	Benchmark dataset with diverse facial expressions
FER Aligned Dataset	Fear (552), Angry (633), Neutral (863), Sadness (904), Happy (1577)	4529	(48, 48, 1)	Aligned dataset for standardized and controlled analysis
FER Aligned + CK Dataset	Surprise (249), Fear (728), Angry (938), Neutral (1230), Sadness (1153), Happy (2203)	6501	(48, 48, 1)	Combined dataset for comprehensive comparative study
RAF-DB Dataset	Surprise (1619), Fear (355), Angry (867), Neutral (3204), Sad (2460), Disgust (877), Happy (5957)	15339	(100, 100, 3)	RGB images with diverse emotions for deep learning models
AffectNet Dataset	Surprise (1851), Fear (1839), Neutral (1880), Sad (1821), Disgust (1740), Contempt (1833), Happy (1862), Anger (1822)	14648	(1024, 1024, 3)	High-resolution images with diverse emotions

B. Image Pre-Processing

Image pre-processing techniques are very useful in enhancing the model generalization and avoid overfitting [19]. One such technique is data augmentation where transformation such as rotation, shifting, shearing, and flipping are applied.

Another pre-processing technique is transforming the image into grayscale. This process entails a transformation of the RGB colour images into black and white with the intensity value is between 0 to 255. Grayscale images simplify the data representation, making it easier for the model to extract relevant features and patterns.

In addition, one-hot encoding is applied where categorical data such as the emotion labels are transformed into binary representation. In this study, we would like to examine how image pre-processing can make a valuable contribution in enhancing the performance and reliability of the selected deep learning models.

C. Transfer Learning

Transfer learning is applied where the selected deep-learning models are pre-trained on ImageNet and reused in FER. Fine-tuning with low learning rate is applied to better customize the DenseNet121, InceptionV3, ResNet50, and Xception.

D. Model Architecture

The selections of DenseNet121, ResNet50, InceptionV3, and Xception for FER are based on their unique architectural innovations that cater to certain issues in deep learning as mentioned in Section I. The architecture of each model is reported as below.

1) DenseNet121: Fig. 1's architecture was proposed by [20], where the dense connectivity pattern leads to a better gradient flow, feature propagation and reuse.

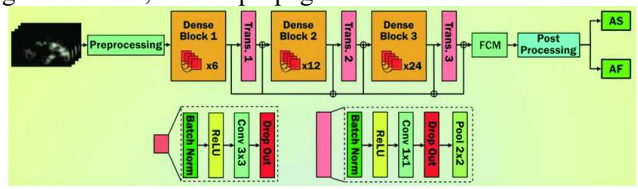

Fig. 1. DenseNet121's architecture. [21]

2) ResNet50: Fig. 2's model was introduced by [22], uses residual learning to address memory consumption in deep networks by learning residual functions. This architecture, comprising multiple layers with residual connections, enables the training of very deep networks.

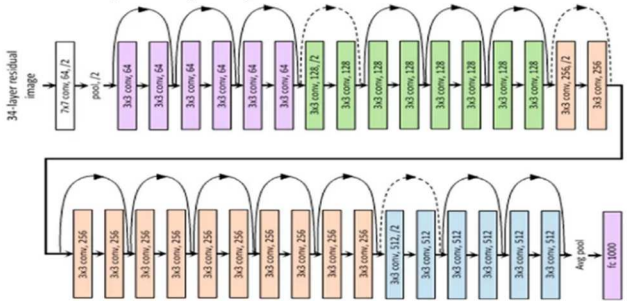

Fig. 2. ResNet50's architecture. [23]

3) InceptionV3: Fig. 3 shows InceptionV3's architecture [24], stands out with its efficient use of the resources and it is one of the first architectures to achieve multi-scale processing. Its architecture is based on stacked layers which includes the occasional max-pooling layer to reduce the grid resolution, as well as dimensional reductions and projections relying on 1x1 convolutions in order to make it computationally efficient.

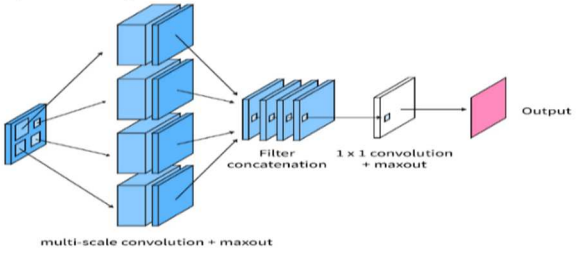

Fig. 3. InceptionV3's architecture. [25]

4) Xception: Fig. 4 (Xception), is an improved version of InceptionV3 which was proposed by [26]. It allows spatial correlation and cross channel correlation separation in feature maps which makes parameter usage and performance more efficient.

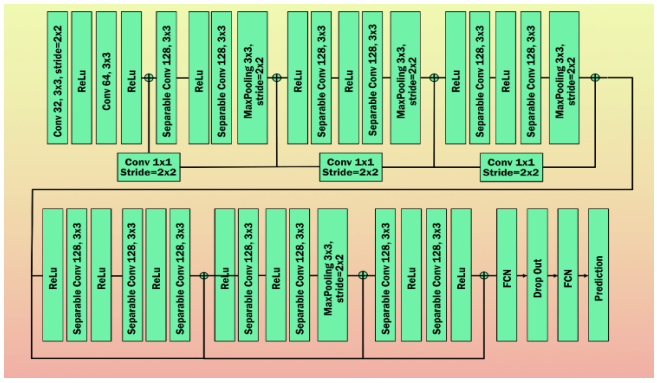

Fig. 4. Xception's architecture. [21]

E. Measurement Metrics

In this study, classification accuracy, recall and test loss are included as the measurement metric. Classification accuracy (refer equation (1)) is a ratio of number correct predictions to the total predictions made by the model. Recall (refer equation (2)), also known as sensitivity or true positive rate, is the ratio of actual positive observations to accurately predicted positive observations, underscoring model's capability to find true positives and avoid false negatives.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Test loss is determined by the categorical cross-entropy loss function and it reveals the discrepancies between the model's predictions and the actual values. Lower test loss means better prediction performance while higher test loss means higher difference between predicted values and actual values.

F. Process Flow

Some of the terminologies used in this study are: models with pre-defined input dimensions are named as custom models (the verification of input size is reported in Section IV.B); for models with transfer learning, it is named as pre-trained models. Pre-processing techniques are applied on all the input data for all the deep learning models in this study. However, for pre-trained model, grayscale transformation is not applied.

Fig. 5 shows the process flow of the custom model for FER. The flow of the custom model encompasses multiple stages, namely input size customization, load dataset, image preprocessing, separate train and test set, custom model, model setup, and evaluation. Fig. 6 depicts the process flow of the pre-trained model for FER. The flow of the pre-trained model includes load dataset, image preprocessing, separate train and test set, pre-trained model, build model on top of pre-trained model, train the top layer, fine-tuning the entire model, and evaluation.

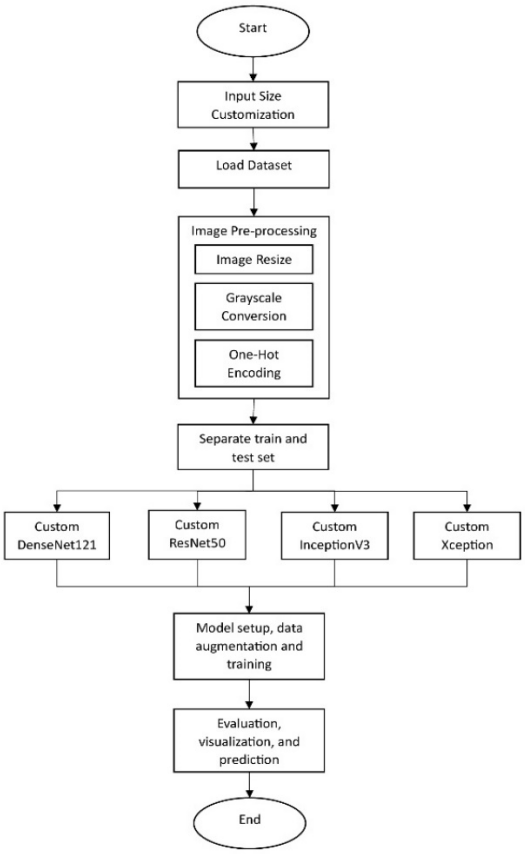

Fig. 5. Process flow of the custom model.

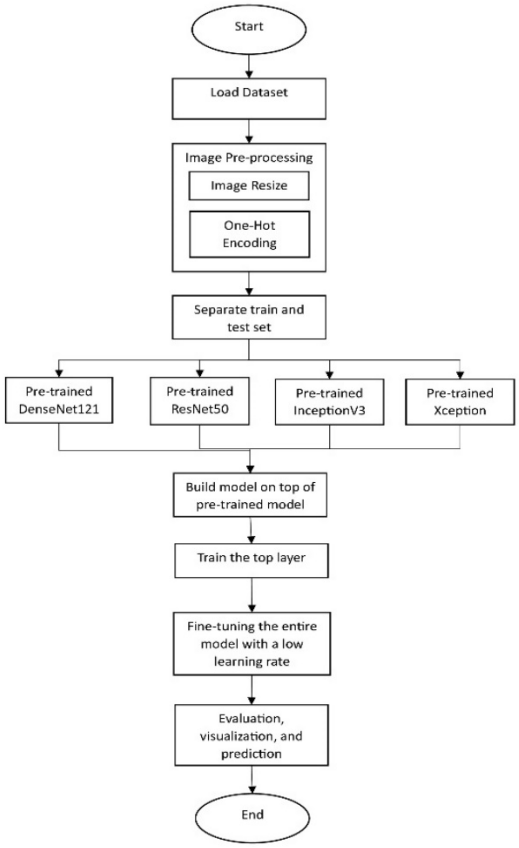

Fig. 6. Process flow of the pre-trained model.

IV. Experimental Setup

In this study, in order to encourage the use of standard platform, a cloud-based system is suggested to apply. The specification of the chosen cloud-based platform in this study consists of 2 NVIDIA Tesla T4 GPUs on Kaggle to enable parallel processing for deep learning tasks. Maximum RAM allowed is 29 GB. An 80-20 ratio of training and testing data is used to increase the accuracy of the model. In this section, the setting of hyperparameter, verification of input size, performance comparison among models which include the sub-class of each emotion are discussed.

A. Hyperparameter

The hyperparameters for data augmentation is applied to all the datasets. These include the rotation range with 8, width shift range was set as 0.08, height shift range was 0.08, shear range was 0.1 and image was horizontal flip.

The hyperparameters for all the custom deep learning models are listed in Table II. All the custom models were set with similar batch size, i.e. 32 and the optimizer was Adam. The hyperparameters for the pre-trained models are listed in Table III with each model was set with similar optimizer, i.e. Adam.

TABLE II. Hyperparameter of Custom Models

	learning rate	epochs
Custom DenseNet121	0.001	100
Custom ResNet50	0.001	100
Custom InceptionV3	0.001	100
Custom Xception	0.01	16

TABLE III. Hyperparameter of Pre-trained Models

	learning rate	epochs	batch size
Pre-trained DenseNet121	0.0001	100	32/64
Pre-trained ResNet50	0.00001	100	32/64
Pre-trained InceptionV3	0.00001	50/100	32
Pre-trained Xception	0.00001	50/100	32

B. Verification of Input Size

In order to identify the custom size of input for custom model, a default CNN architecture with Conv2D, MaxPooling2D, Flatten, and Dense layers (ReLU and softmax activations) was implemented. Different image resolutions (48x48, 64x64, 128x128, and 256x256) were tested to evaluate the impact of different image resolution on the classification accuracy and test loss of the selected deep learning models across various datasets (CK, FER aligned, FER aligned + CK, RAF-DB, and AffectNet).

Theoretically, we assumed that lower image resolutions (48x48 and 64x64) use fewer computational resources but capture less detail with lower classification accuracy, while higher image resolutions (128x128 and 256x256) offer more detail with higher classification accuracy but require more computational power. However, the experimental results show that higher image resolution not necessary contribute to better classification accuracy and lower test loss. In Table IV, the results with higher classification accuracy (denoted as Acc. In Table IV) but lower test loss was bolded. Majority datasets obtain good classification accuracy with low test loss when the image resolution is 48x48 (i.e. FER aligned and RAF-DB) or 64x64 (i.e. CK, FER aligned + CK).

TABLE IV. TEST SET ACCURACIES FOR DIFFERENT INPUT IMAGE SIZES

	Input Image Sizes			
	48x48	*64x64*	*128x128*	*256x256*
CK (Acc.)	0.9933	**1**	1	1
CK (Test Loss)	0.0688	**0.0210**	0.0291	0.0279
FER aligned (Acc.)	**0.6689**	0.6611	0.5993	0.5762
FER aligned (Test Loss)	**0.9445**	1.0288	1.4704	1.6840
FER aligned + CK (Acc.)	0.6987	**0.7048**	0.6749	0.6426
FER aligned + CK (Test Loss)	0.9238	**0.9513**	1.4766	1.2126
RAF-DB (Acc.)	**0.6640**	0.6571	0.6359	0.5242
RAF-DB (Test Loss)	**1.5082**	1.8704	2.2604	3.8042
AffectNet (Acc.)	0.5205	0.5413	**0.5809**	0.5652
AffectNet (Test Loss)	1.3820	1.5507	**2.3321**	2.9979

Based on the preliminary results in Table IV, we listed the selected input size for each dataset of custom model in Table V.

TABLE V. SELECTION OF CUSTOM MODEL'S INPUT IMAGE SIZES FOR FER DATASETS

Datasets	Input Image Sizes	Reasoning
CK	64x64	Good accuracy in input image size experiments, highest average accuracy across sizes
FER aligned	48x48, 64x64	Good accuracy for 48x48, highest average accuracy with 64x64
FER aligned + CK	64x64	Good accuracy in input image size experiments, highest average accuracy across sizes
RAF-DB	48x48, 64x64	Good accuracy for 48x48, highest average accuracy with 64x64
AffectNet	128x128, 64x64	Good accuracy for 128x128, highest average accuracy with 64x64

For pre-trained models, we do not verify the input size through CNN but we use the selected models and run on few input sizes. The experiment was started from the minimum allowed input sizes for each model and varying the sizes till the default input size. Note that the minimum allowed sizes for DenseNet121 and ResNet50 are 32x32, for InceptionV3 is 75x75 and Xception is 71x71. For larger default sizes are: 224x224 for DenseNet121 and ResNet50; and 299x299 for InceptionV3 and Xception. Table VI shows the summary of various input sizes for each pre-trained model.

TABLE VI. VAROUS INPUT SIZES FOR PRE-TRAINED MODEL

Pre-trained Model	Minimum Input Size	Default Input Size	Others Input Sizes
DenseNet121	32x32	224x224	71x71, 75x75, 128x128, 299x299
ResNet50	32x32	224x224	71x71, 75x75, 128x128, 299x299
InceptionV3	75x75	299x299	128x128, 224x224
Xception	71x71	299x299	75x75, 128x128, 224x224

V. COMPARISON AMONGST MODELS

The comparison among models in this section provides a comprehensive analysis of custom and pre-trained deep learning models for FER across various datasets. Through accuracy assessments and recall analysis, the impact of preprocessing techniques, transfer learning, and model architectures on classification performance is evaluated.

A. Results of all Models

Table VII shows the accuracy results of various deep learning algorithms on different datasets. For pre-trained models, input size with the highest testing accuracy are selected. The input size was listed under the column 'Models'. For CK dataset, Custom ResNet50 and Custom Xception achieved 100% testing accuracy. Pre-trained DenseNet121 yielded a higher testing accuracy of 0.7318, demonstrating its ability to capture complex features for emotion recognition in FER aligned. In the FER aligned + CK dataset, Pre-trained DenseNet121 again achieved the highest testing accuracy, i.e. 0.7763. The strength of DenseNet121 is also shown in RAF-DB dataset, where the image input was maintained in RGB mode. Pre-trained DenseNet121 achieved 0.8488 testing accuracy. Lastly, on the AffectNet dataset, Pre-trained DenseNet121 again scored the highest with 0.6304 testing accuracy, highlighting its capability to recognize subtle facial expressions.

For the impact of input size among pre-trained models, it seems like consistency of image resolution depends on the type of deep learning model. It is found that DenseNet121 can well classified all the datasets with 128x128 and 224x224; InceptionV3 with 128x128. and Xception mostly classified well when images are having 224x224. However, different best classification accuracy was obtained with different input size for ResNet50.

Then, the classification strength between custom models and pre-trained models were compared and presented in Fig. 7. Among all the models, DenseNet121 has the highest classification accuracy, i.e. 0.7908. The dense connectivity in DenseNet121 and feature reuse are well contributed to its efficient learning, making it suitable for various datasets targeting the FER. In terms of custom model, ResNet50 has the best achievement among all the compared models, i.e. 0.7651.

In general, different input sizes do impact on the model performance. However, although higher dimensions can have more image details but it does not help in the classification task by deep learning algorithms. In other words, not necessarily higher dimensions are equivalent to higher classification accuracy. Besides that, the results in this section show that transfer learning with pre-trained weights can enhance the model's generalization ability and classification performance.

TABLE VII. ACCURACY OF ALL MODELS

Datasets	Deep Learning	Models	Training Accuracy	Testing Accuracy
CK	DenseNet121	Custom (64x64)	0.9283	0.8867
		Pre-trained (128x128)	0.9717	0.9667
	ResNet50	**Custom (64x64)**	**1.0000**	**1.0000**
		Pre-trained (128x128)	0.9483	0.9467
	InceptionV3	Custom (64x64)	0.9783	0.9667
		Pre-trained (224x224)	0.9583	0.9200

	Xception	**Custom (64x64)**	**1.0000**	**1.0000**
		Pre-trained (224x224)	0.9600	0.9533
FER aligned	**DenseNet121**	Custom (48x48)	0.8173	0.6998
		Pre-trained (224x224)	**0.9008**	**0.7318**
	ResNet50	Custom (64x64)	0.8587	0.7053
		Pre-trained (75x75)	0.7221	0.6600
	InceptionV3	Custom (48x48)	0.7750	0.6843
		Pre-trained (224x224)	0.6583	0.5828
	Xception	Custom (48x48)	0.7767	0.6810
		Pre-trained (224x224)	0.8460	0.6876
FER aligned + CK	**DenseNet121**	Custom (64x64)	0.8079	0.7440
		Pre-trained (128x128)	**0.9792**	**0.7763**
	ResNet50	Custom (64x64)	0.8690	0.7594
		Pre-trained (75x75)	0.7054	0.6956
	InceptionV3	Custom (64x64)	0.8398	0.7502
		Pre-trained (128x128)	0.7694	0.6733
	Xception	Custom (64x64)	0.8365	0.7663
		Pre-trained (224x224)	0.7500	0.7171
RAF-DB	**DenseNet121**	Custom (64x64)	0.7792	0.7718
		Pre-trained (128x128)	**0.9913**	**0.8488**
	ResNet50	Custom (64x64)	0.8919	0.8227
		Pre-trained (71x71)	0.8435	0.7999
	InceptionV3	Custom (64x64)	0.9128	0.7956
		Pre-trained (128x128)	0.8922	0.7735
	Xception	Custom (64x64)	0.8075	0.7930
		Pre-trained (128x128)	0.9333	0.7917
AffectNet	**DenseNet121**	Custom (128x128)	0.5847	0.5338
		Pre-trained (128x128)	**0.9050**	**0.6304**
	ResNet50	Custom (128x128)	0.6481	0.5379
		Pre-trained (128x128)	0.6306	0.5307
	InceptionV3	Custom (64x64)	0.4866	0.4618
		Pre-trained (128x128)	0.7318	0.5505
	Xception	Custom (64x64)	0.5533	0.5225
		Pre-trained (75x75)	0.9088	0.5631

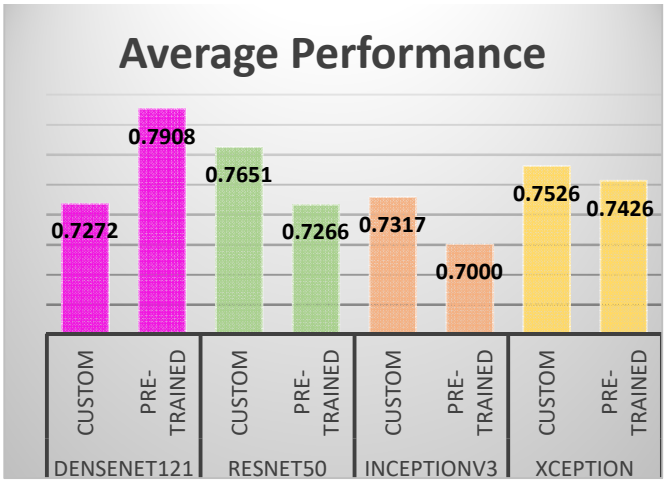

Fig. 7. Average performance of deep learning models.

B. Results Comparison with Various Emotion

In this section, classification accuracy and recall are analysed for each emotion. The models used in this experiment is similar to Section A. Table VIII lists the index of each model used in the experiments to ease the presentation from Tables IX to XIII. Index 'a' represents the custom model whereas index 'b' represents the pre-trained model.

TABLE VIII. EXPERIMENT MODELS

Experiments	Model
1a	Custom DenseNet121
1b	Pre-trained DenseNet121
2a	Custom ResNet50
2b	Pre-trained ResNet50
3a	Custom InceptionV3
3b	Pre-trained InceptionV3
4a	Custom Xception
4b	Pre-trained Xception

The average classification performance of each emotion is considered very good (with minimum 84% and maximum 100%) in CK dataset (refer Table IX). Similar to Table VII, Custom ResNet50 and Custom Xception consistently obtained 100% in each emotion. Inconsistency of performance can be observed in Custom DenseNet121 (standard deviation is 27.12), Pre-trained DenseNet121 (standard deviation is 10.55) and Pre-trained InceptionV3 (standard deviation is 16.02) in relation to these emotions: "Fear" and "Sadness". However, majority of the models can well classify CK and this may indicate that CK dataset is an easy and simple dataset.

For FER-aligned dataset (refer Table X), the highest recall rate was shown under "Sadness", while "Neutral" had the lowest recall rate at 53.88% and followed by "Angry" with average 55.38%, "Happy" with 61.88% and "Fear" with 64.25%. The complexity of facial expressions associated with each emotion (except "Sadness") may have influenced these recall rates and causing inconsistencies in the model performance. The standard deviation is ranged from the lowest with Custom ResNet50 (i.e. 11.17) to the highest Pre-trained InceptionV3 (i.e. 16.65).

When both CK and FER aligned are combined together (refer Table XI), the classification performance did improve for "Angry" (in CK, it was 98.50% but now it was 99.25%) and "Neural" (from 53.88% to 68.88%) when comparing the emotion result in Tables IX and X. However, the results

degraded for the other emotions and the range for standard deviation was between 14.79 (Pre-trained DenseNet121) to 23.57 (Pre-trained InceptionV3). The complicated facial expressions within this dataset may led to varying recall rates. Amongst the models, Pre-trained DenseNet121 achieved the highest recall rate with 77.33% across all emotion with the lowest standard deviation 14.79.

The results of RAF-DB (refer Table XII) reflect a bigger variation in recall rate where the minimum was found as 9.00% ("Sad" by custom DenseNet121) and the maximum was 93.00% ("Surprise" by Custom ResNet50). Custom DenseNet121 has the highest standard deviation, i.e. 32.78. Based on the average recall rate, both "Surprise" and "Happy" scored the highest with 90.00% and 81.00% respectively. For other emotions, by sorting the average recall rate from the lowest to the highest is "Sad" (37.00%), "Disgust" (47.00%), "Fear" (68.13%) and "Neutral" (77.75%). Again, Pre-trained DenseNet121 achieved the highest recall rate with 76.14% across all emotion with the lowest standard deviation 13.97.

Amongst all the datasets, AffectNet is having the highest complexity and it contains 8 different emotions (refer Table XIII). The highest recall rates were shown under "Disgust" and "Sadness" with 82.25% and 86.25%, respectively. Four emotions scored the recall rates below 50.00%, i.e. "Fear" with the lowest at 35.38%, followed by "Happy" with 40.00%, "Angry" with 40.38%, "Surprise" with 41.75% and "Contempt" with 45.38%. The average recall for "Neutral" is 69.47%. The deep learning model with the smallest standard deviation was Pre-Trained Xception but the recall rate was the second highest, i.e. 56.00%. Pre-trained DenseNet121 has the highest recall rate, i.e. 63.00% but the standard deviation is slightly higher than Pre-Trained Xception, i.e. 16.81.

TABLE IX. STATISTICAL ANALYSIS ON THE SENSITIVITIES (RECALL) ON CK DATASET

Exp	Angry	Fear	Happy	Sadness	Surprise
1a	98.00	47.00	100.00	53.00	100.00
1b	98.00	100.00	100.00	76.00	100.00
2a	100.00	100.00	100.00	100.00	100.00
2b	98.00	87.00	85.00	94.00	100.00
3a	98.00	93.00	89.00	100.00	100.00
3b	100.00	67.00	93.00	71.00	100.00
4a	100.00	100.00	100.00	100.00	100.00
4b	96.00	80.00	96.00	94.00	100.00
Avg	**98.50**	**84.25**	**95.38**	**84.75**	**100.00**

a. All values are in %

TABLE X. STATISTICAL ANALYSIS ON THE SENSITIVITIES (RECALL) ON FER ALIGNED DATASET

Exp	Angry	Fear	Happy	Neutral	Sadness
1a	52.00	72.00	69.00	55.00	84.00
1b	63.00	71.00	66.00	55.00	92.00
2a	59.00	68.00	64.00	59.00	86.00
2b	57.00	58.00	59.00	47.00	87.00
3a	56.00	69.00	63.00	56.00	83.00
3b	43.00	43.00	51.00	45.00	82.00
4a	59.00	68.00	63.00	50.00	85.00
4b	50.00	65.00	60.00	64.00	85.00
Avg	**55.38**	**64.25**	**61.88**	**53.88**	**85.50**

b. All values are in %

TABLE XI. STATISTICAL ANALYSIS ON THE SENSITIVITIES (RECALL) ON FER ALIGNED + CK DATASET

Exp	Angry	Fear	Happy	Neutral	Sadness	Surprise
1a	100.00	57.00	74.00	70.00	55.00	90.00
1b	100.00	63.00	73.00	71.00	66.00	91.00
2a	100.00	59.00	74.00	70.00	61.00	91.00
2b	98.00	53.00	69.00	64.00	46.00	87.00
3a	98.00	56.00	73.00	68.00	60.00	91.00
3b	100.00	45.00	61.00	67.00	40.00	88.00
4a	98.00	59.00	73.00	74.00	58.00	92.00
4b	100.00	48.00	71.00	67.00	60.00	88.00
Avg	**99.25**	**55.00**	**71.00**	**68.88**	**55.75**	**89.75**

c. All values are in %

TABLE XII. STATISTICAL ANALYSIS ON THE SENSITIVITIES (RECALL) ON RAF-DB DATASET

Exp	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
1a	78.00	16.00	67.00	83.00	73.00	9.00	89.00
1b	85.00	62.00	70.00	84.00	84.00	55.00	93.00
2a	82.00	55.00	79.00	80.00	80.00	39.00	93.00
2b	78.00	57.00	71.00	79.00	76.00	34.00	92.00
3a	73.00	40.00	49.00	88.00	80.00	76.00	82.00
3b	75.00	50.00	66.00	76.00	72.00	25.00	91.00
4a	78.00	45.00	75.00	80.00	78.00	24.00	90.00
4b	77.00	51.00	68.00	78.00	79.00	34.00	90.00
Avg	**78.25**	**47.00**	**68.13**	**81.00**	**77.75**	**37.00**	**90.00**

d. All values are in %

TABLE XIII. STATISTICAL ANALYSIS ON THE SENSITIVITIES (RECALL) ON AFFECTNET DATASET

Exp	Angry	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
1a	44.00	35.00	87.00	35.00	36.00	62.00	88.00	37.00
1b	48.00	52.00	90.00	51.00	56.00	65.00	88.00	54.00
2a	39.00	44.00	82.00	41.00	41.00	58.00	86.00	39.00
2b	42.00	39.00	80.00	43.00	40.00	57.00	84.00	39.00
3a	26.00	48.00	77.00	7.00	11.00	62.00	90.00	45.00
3b	39.00	45.00	86.00	42.00	44.00	57.00	87.00	37.00
4a	40.00	48.00	81.00	23.00	40.00	60.00	87.00	37.00
4b	45.00	52.00	75.00	41.00	52.00	57.00	80.00	46.00
Avg	**40.38**	**45.38**	**82.25**	**35.38**	**40.00**	**69.47**	**86.25**	**41.75**

e. All values are in %

VI. CONCLUSION

In this comparative study, a diverse range of FER datasets including CK, FER, FER aligned + CK, RAF-DB, and AffectNet were utilized to evaluate and access the performance of the deep learning algorithms. Techniques such as data augmentation, grayscale conversion, and one-hot encoding were applied for both custom and pre-trained models. The impact of transfer learning and input dimension were studied. As a conclusion, the results showed that not necessarily higher input dimension will have better classification performance. Besides that, transfer learning does significantly improve the performance and generalization of deep learning models. Notably, Pre-trained DenseNet121 demonstrated the best classification accuracy and recall across all datasets. Moreover, the value of recall metric provided valuable insights into the models' abilities to correctly identify each emotion. When evaluating each

emotion with the recall, some emotions can be predicted well by the deep learning models but some emotions are not. The only exception is CK dataset and it obtains a consistency of classification performance among all emotions and among all deep learning models.

Thus, we suggest that researchers should include more evaluation metric in identifying the performance of their proposed model in each emotion. As shown in this study, more insights information can be seen and explored. Then, a new FER algorithm that assess with only CK dataset is not enough. More complicated FER datasets such as FER aligned, RAF-DB and AffectNet should be utilized to evaluate the performance of proposed models. Besides that, transfer learning seems like can improve the adaptability and generalizability when dealing with different datasets. However, the experimental results show that the good performance may not be true for other deep learning algorithms although transfer learning is applied. We encourage the code released among researchers to ease the benchmarking and comparison among different FER algorithms. Here is the released code of this comparative study that is available at GitHub: https://github.com/JoskenGan/A-Comparative-Study-On-Facial-Emotion-Recognition

In future, the impact of each different pre-processing techniques and sample size of each emotion can be further explored. At the same time, the execution time should be recorded to know the efficiency of the deep learning algorithms.

ACKNOWLEDGMENT

The research appreciates Multimedia University for supporting the work with the IRFund (grant number is MMUI/210025).

REFERENCES

[1] I. Kosunen, "*Facial Expressions in Human-Computer Interaction (HCI)*," www.linkedin.com. https://www.linkedin.com/pulse/facial-expressions-human-computer-interaction-hci-ilkka-kosunen-kq4cc (accessed Jul. 20, 2024).

[2] M. Leo, P. Carcagnì, P. L. Mazzeo, P. Spagnolo, D. Cazzato, and C. Distante, "Analysis of Facial Information for Healthcare Applications: A survey on Computer Vision-Based Approaches," *Information*, vol. 11, no. 3, p. 128, Feb. 2020, doi: 10.3390/info11030128.

[3] S. Cosentino, E. Randria, J. Lin, T. Pellegrini, S. Sessa, and A. Takanishi, "Group emotion recognition Strategies for entertainment robots," *HAL Open Science*, pp. 813–818, Oct. 2018, doi: 10.1109/iros.2018.8593503.

[4] A. Stahelski, A. Anderson, N. Browitt, and M. Radeke, "Facial expressions and emotion labels are separate initiators of trait inferences from the face," *Frontiers in Psychology*, vol. 12, Dec. 2021. doi:10.3389/fpsyg.2021.749933.

[5] C. M. Klingner and O. Guntinas-Lichius, "Mimik und emotion," *Laryngo-Rhino-Otologie*, vol. 102, no. S 01, May 2023. doi:10.1055/a-2003-5687.

[6] S. M. Mavadati, M. H. Mahoor, K. T. Bartlett, P. T. Trinh, and J. F. Cohn, "DISFA: a Spontaneous Facial Action Intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/t-affc.2013.4.

[7] A.-L. Cîrneanu, D. Popescu, and D. D. Iordache, "New Trends in Emotion Recognition using Image Analysis by Neural Networks, A Systematic review," *Sensors*, vol. 23, no. 16, p. 7092, Aug. 2023, doi: 10.3390/s23167092.

[8] M. Sutar and A. Ambhaikar, "A Comparative Study on Deep Facial Expression Recognition," *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2023, doi: 10.1109/iciccs56967.2023.10142703.

[9] A. Abbas and S. K. Chalup, "The Impact of Image Resolution on Facial Expression Analysis with CNNs," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, Jul. 2019, doi: 10.1109/ijcnn.2019.8852264.

[10] S. R. Sekaran, C. P. Lee, and K. M. Lim, "Facial emotion recognition using transfer learning of AlexNet," *2021 9th International Conference on Information and Communication Technology (ICoICT)*, Aug. 2021, doi: 10.1109/icoict52021.2021.9527512.

[11] T. Teixeira, É. Granger, and A. L. Koerich, "Continuous Emotion Recognition with Spatiotemporal Convolutional Neural Networks," *Applied Sciences*, vol. 11, no. 24, p. 11738, Dec. 2021, doi: 10.3390/app112411738.

[12] S. Gupta, P. Kumar, and R. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11365–11394, Sep. 2022, doi: 10.1007/s11042-022-13558-9.

[13] M. Bentoumi, M. Daoud, M. Benaouali, and A. T. Ahmed, "Improvement of emotion recognition from facial images using deep learning and early stopping cross validation," Multimedia Tools and Applications, vol. 81, no. 21, pp. 29887–29917, Apr. 2022, doi: 10.1007/s11042-022-12058-0.

[14] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021, doi: 10.3390/s21093046.

[15] A. John, A. Mc, A. S. Ajayan, S. Sanoop, and V. R. Kumar, "Real-Time facial emotion recognition system with improved preprocessing and feature extraction," *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Aug. 2020, doi: 10.1109/icssit48917.2020.9214207.

[16] J A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive Correlation-Based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26756–26768, Jan. 2022, doi: 10.1109/access.2022.3156598.

[17] W. Hayale, P. Negi, and M. H. Mahoor, "Facial Expression Recognition Using Deep Siamese Neural Networks with a Supervised Loss function," 2019 14th IEEE International Conference on Automatic Face &Amp; Gesture Recognition (FG 2019), May 2019, doi: 10.1109/fg.2019.8756571.

[18] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, "A survey of AI-based facial emotion recognition: Features, ML & DL techniques, age-wise datasets and Future Directions," *IEEE Access*, vol. 9, pp. 165806–165840, 2021. doi:10.1109/access.2021.3131733.

[19] A. L. Pereira, L. A. Fernandes, and A. Conci, "Image preprocessing techniques for facial expression classification," *Proceedings of the 2nd Life Improvement in Quality by Ubiquitous Experiences Workshop (LIQUE 2022)*, Jun. 2022. doi:10.5753/lique.2022.19995.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708, Jan.2018.

[21] S. S. Sanagala, A. Nicolaides, S. K. Gupta, V. K. Koppula, L. Saba, S. Agarwal, A. M. Johri, M. S. Kalra, and J. S. Suri, "Ten Fast Transfer Learning Models for Carotid Ultrasound Plaque Tissue Characterization in Augmentation Framework Embedded with Heatmaps for Stroke Risk Stratification," *Diagnostics*, vol. 11, no. 11, p. 2109, Nov. 2021, doi: 10.3390/diagnostics11112109.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, doi: 10.1109/cvpr.2016.90.

[23] S. Bangar, "*Resnet Architecture Explained*," Medium. https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d (accessed Jul. 20, 2024).

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, doi: 10.1109/cvpr.2015.7298594.

[25] D. Sivakumar, "*Introduction to InceptionNet*," Scaler Topics. https://www.scaler.com/topics/inception-network/ (accessed Jul. 20, 2024).

[26] F. Chollet, "XCeption: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, doi: 10.1109/cvpr.2017.195