International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 3
ISSN: 2088-5334

Comparison and Analysis of CNN Models to Improve a Facial Emotion Classification Accuracy for Koreans and East Asians

Jun-Hyeong Lee [a], Ki-Sang Song [a,*]

[a] *Computer Education, Korea National University Education, 28173, Republic of Korea*
Corresponding author: [*]kssong@knue.ac.kr

Abstract **- Facial emotion recognition is one of the popular tasks in computer vision. Face recognition techniques based on deep learning can provide the best face recognition performance, but using these techniques requires a lot of labeled face data. Available large-scale facial datasets are predominantly Western and contain very few Asians. We found that models trained using these datasets were less accurate at identifying Asians than Westerners. Therefore, to increase the accuracy of Asians' facial identification, we compared and analyzed various CNN models that had been previously studied. We also added Asian faces and face data in realistic situations to the existing dataset and compared the results. As a result of model comparison, VGG16 and Xception models showed high prediction rates for facial emotion recognition in this study. and the more diverse the dataset, the higher the prediction rate. The prediction rate of the East Asian dataset for the model trained on FER2013 was relatively low. However, for data learned with KFE, the model prediction of FER2013 was predicted to be relatively high. However, because the number of East Asian datasets is small, caution is needed in interpretation. Through this study, it was confirmed that large CNN models can be used for facial emotion analysis, but that selection of an appropriate model is essential. In addition, it was confirmed once again that a variety of datasets and the prediction rate increase as a large amount of data is learned.**

***Keywords*— Computer vision; AI; CNN; facial classification; facial emotion recognition.**

Manuscript received 30 Aug. 2023; revised 9 Oct. 2023; accepted 22 Feb. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

Facial emotion recognition (FER) is one of the significant tasks in computer vision. In classifying these emotions, Ekman and Friesen [1] have classified human facial emotions into seven types: anger, contempt, fear, joy, normal, sadness, and surprise. This is one of the leading indicators for classifying human facial emotions in computer vision. Recently, computer vision has made significant progress due to the development of the Convolutional Neural Network (CNN) [2]. Generally, CNN-based tasks have two essential elements: the structure of the CNN and the training data set. Diverse face datasets are essential to advance FER research. Still, most public face datasets primarily consist of Western face images and contain only a small number of East Asian face images. Therefore, in the case of Asian face recognition, we found that deep learning models trained with these data sets provide lower face recognition accuracy than Western face recognition rates.

In addition, FER classification shows significant differences between individuals and in manipulated and wild environments. In the case of inter-individual differences, an individual's face varies depending on gender, age, or racial group. Changes in the subject include facial color, lighting, and head posture variations. Despite these challenges, research on FER has attracted much interest and has led to several practical applications in human-computer interaction systems and data analysis [3].

Accordingly, one of the methods for finding faces is a landmark-based algorithm. One of the most recent papers, a study by Li [4], uses an adaptive feature fusion network to recognize faces. Facial landmark detection can achieve surprising results under controlled conditions in the laboratory. However, in noisy (wild) environments, they generally do not work well due to changes in head posture, lighting, etc. In recent research, attention mechanisms for image classification problems have been developed to increase the performance of CNN by focusing on small details [5]. Moreover, in image segmentation problems, CNN effectively derives valid data by searching pixel units in images and classifying them into practical semantic units [6].

To solve this problem, it was confirmed that it was necessary to add an Asian face dataset by referring to various previous studies. In addition to adding Asian datasets, this study uses previously studied datasets and CNN models to find a model with a high facial emotion recognition rate for Asians. It compares models based on dataset learning to determine the differences.

II. Material and Method

Comparison of CNN models require the same dataset and various CNN models. This study used the most general-purpose dataset and CNN model with recognized operation characteristics. The public datasets and CNN models used in this study are as follows.

A. Data Set

FER2013 [7] is the most basic data set for facial emotion classification. There are 7 types of facial emotions. The files are extracted separately from the person's face and converted to gray with a size of 48*48. It is structured so that it can be used for various machine learning. It is the most basic data set in facial emotion classification research. In this study, 28,709 of the most widely used learning sheets were used for learning. For verification purposes, 3,589 public pages were used.

Fig. 1 Sample of facial emotion image data from FER2013

CK+ [8] is a dataset containing 593 video sequences from 123 subjects ranging from 18 to 50 years of age and of various genders and races. Each video shows facial changes from neutral expressions recorded at 30 frames per second (FPS) with a resolution of 640x490 or 640x480 pixels. Of these videos, 327 were categorized into one of seven expression types: anger, contempt, disgust, fear, happiness, sadness, and surprise. In this study, preprocessed facial photos of 48*48 were used. A total of 902 images were used.

Fig. 2 Sample of facial emotion image data from CK+

Korean Facial Emotion (KFE) is data preprocessed for artificial intelligence learning by downloading facial emotion data from the DB website AI Hub [9] provided by Korea's Korea Intelligence and Information Society Promotion Agency. A separate dataset was extracted for this study. Only the face is extracted from the downloaded image using the Haarcascade [10] face extraction function. Afterward, it is converted to a gray 48*48 size. Since Haarcascade's face extraction function was inaccurate, a separate expert performed facial emotion labeling to remove images that were not human faces or were incorrect. A total of 3,082 pieces of facial emotion data were used in this study. Preprocessing was performed appropriately for artificial intelligence learning. Likewise, pictures of the emotions on Korean faces are provided. It was pre-processed and converted into a gray image of size 48*48.

Fig. 3 Sample of facial emotion image data from KFE

JAFFE [11] is a data set containing facial emotional expressions of Japanese women. Miyuki Kamachi and Michael J. Lyons collected images of facial expressions containing a variety of emotions from Japanese women and made them available for research and experimentation. Seven facial emotions from about 10 Japanese women were collected. A total of 213 images were used. The photos were taken from the front with the face removed from the jewelry.

Fig. 4 Sample of facial emotion image data from JAFFE

B. Method

In most traditional approaches, the actual first step is to detect the position of the face and then extract geometric features, shape features, or both to generate specific vectors for the model. These methods are usually very complex and require many technical manipulations. When data becomes large, characterization becomes very difficult. These methods often need help in natural or noisy environments where landmark detection is difficult. In the face of these difficulties, deep learning's CNN has solved the problem.

CNN is a field of artificial intelligence and machine learning. It is one of the neural network structures widely used in machine learning and deep learning. It is mainly suitable for computer vision tasks such as image and pattern recognition.

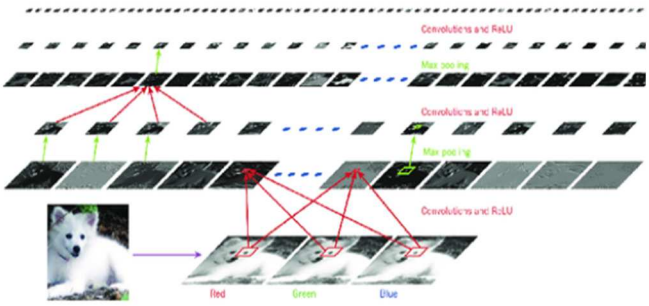

Fig. 5 Architecture of CNN [2]

CNN's structure consists of a convolution layer, pooling layer, activation function, and fully connected layer. CNN is used in various fields, such as computer vision, natural language processing, and speech processing. There are many different CNN models in artificial intelligence for image processing.

There are some representative types of modern neural networks for image classification among various CNN models, such as DeepLab [12], EfficientNet [13], BoT [14], AutoAugment [15], MobileNetV3 [16], MixNet [17], YOLOv4 [18], Vision Transformers [19], CoAtNet-7 [20],

ViT-e [21], BASIC-L [22] and OmniVec [23]. We selected five models recognized for their performance and prediction rate and conducted a study. In this study, a representative CNN model for facial emotion classification was used as follows.

MobileNet V2 [24] is a lightweight deep neural network architecture developed by Google. It is designed to run efficiently even in resource-constrained environments such as mobile and embedded devices. This architecture deploys deep learning models for computer vision tasks and is particularly suitable for tasks such as real-time object detection and image classification.

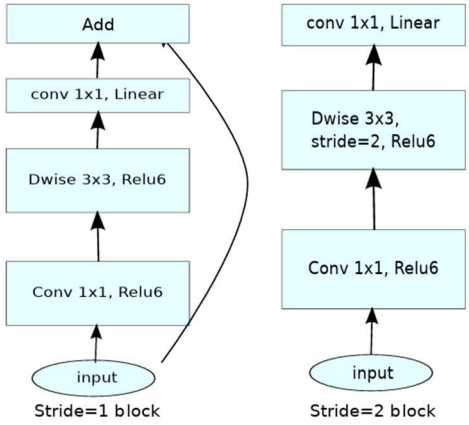

Fig. 6 Architecture of Mobilenet v2

VGG16 [25] consists of sixteen basic layers, including convolutional and fully connected layers. The convolution layer extracts the features of the image, and the fully connected layer classifies the image based on those features. The model was set up with SoftMax and ReLU activation functions. The ReLU activation function filters out negative values and only passes non-negative values to the next layer.

Fig. 7 Architecture of VGG16

Resnet50 [26] comprises one of the deep learning models developed by Microsoft Research. It is a variation of Residual Network (ResNet). It is a neural network consisting of 50 layers. This model is based on the ResNet architecture and has several improvements. 50 deep layers extract detailed image features. It was designed based on residual learning. In addition, information loss was prevented by applying the exact size of convolution in CNN.

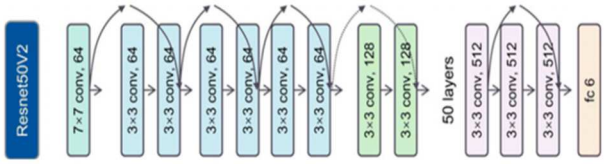

Fig. 8 Architecture of Resnet50v2

Xception [27] is used for image recognition and classification tasks using profound networks in Convolutional Neural Networks. Unlike the existing Inception model, Xception designs a network using depth-wise separable convolution. This makes the convolution operation of the Inception model more efficient, allowing it to achieve high performance. Depth-wise decomposition processes the convolution operation by dividing it into two steps. First, a depth-wise filter is applied to each input channel, followed by a linear combination between the channels via pointwise convolution.

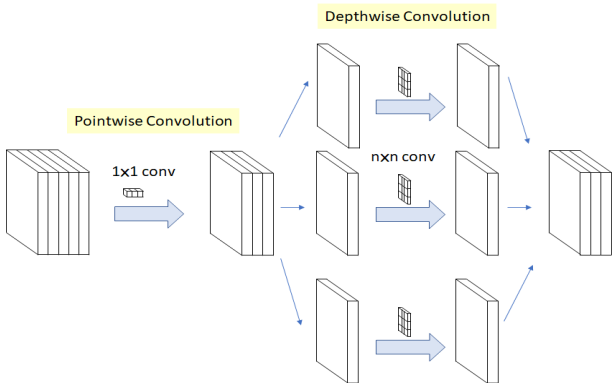

Fig. 9 Architecture of Xception

DenseNet [28] is a neural network architecture based on dense connectivity. This architecture is based on a concept similar to ResNet (Residual Network) but has a structure in which all layers are more closely connected. Because of this, DenseNet alleviates the gradient loss problem and allows the construction of a deeper neural network while reducing the number of model parameters through efficient parameter sharing. Densenet-121 consists of 5 blocks and 3 levels, all of which are interconnected. Each block consists of several convolutional layers and max pooling layers. After block concatenation and flattening, the output matrix consists of two layers for fully connected binary classification.

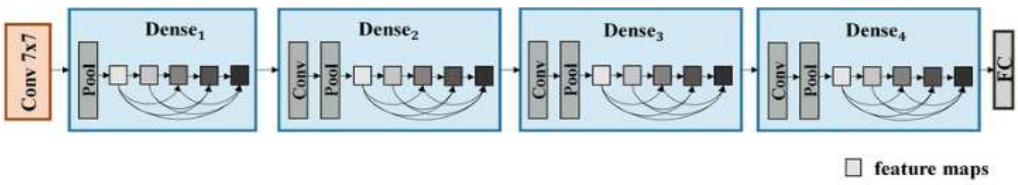

Fig. 10 Architecture of DenseNet [2]

C. Related Works

ImageNet [29] is a dataset that serves as a reference point for evaluating the performance of CNN models. Created in 2009, it contains over 20,000 labels and over 14 million images. Each CNN model has been developed to classify these images accurately. The image classification prediction

rate for ImageNet for each model is as follows. This helps predict the performance of each model.

TABLE I
IMAGENET PREDICTION RATE FOR EACH MODEL [30]

Model	accuracy
VGG-16	0.715
MobileNetV2	0.728
ResNet50	0.761
DenseNet 121	0.770
Xception	0.790

The CNN model was compared with the five models presented above. To classify each model into the same seven classes, a separate layer was added to the final output. Max pooling, global average pooling, and Dense were applied to the added layer so as not to modify the existing extracted feature values as much as possible. The detailed layers for each model are as shown in the following figures.

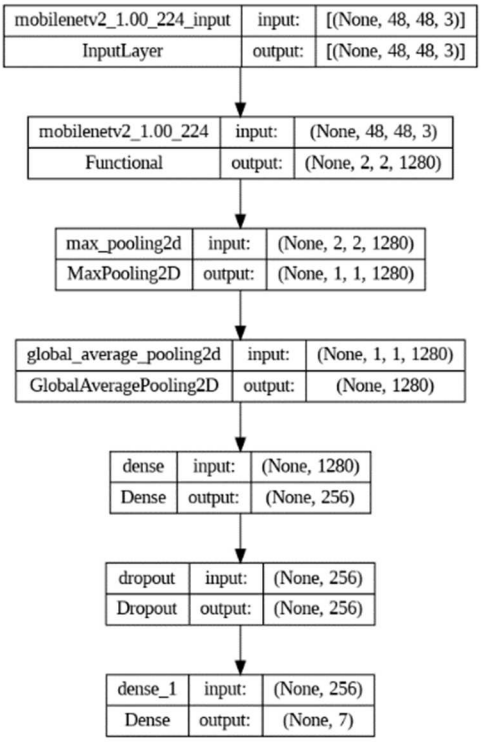

Fig. 11 FER Mobile net architecture modified for FER

The CNN model was compared with the five models presented above. A separate layer was added to the final output to classify each model into the same seven classes. Max pooling, global average pooling, and Dense were applied to the added layer not to modify the existing extracted feature values as much as possible. Max pooling is dividing input data into small areas, extracting the maximum value from each location, and outputting it. It is used to reduce spatial dimensions and emphasize features. Global average pooling refers to averaging all values in each feature map and compressing them into one value. It is usually used in the last layer of a CNN and generates the final output by summarizing valuable information from each feature map.

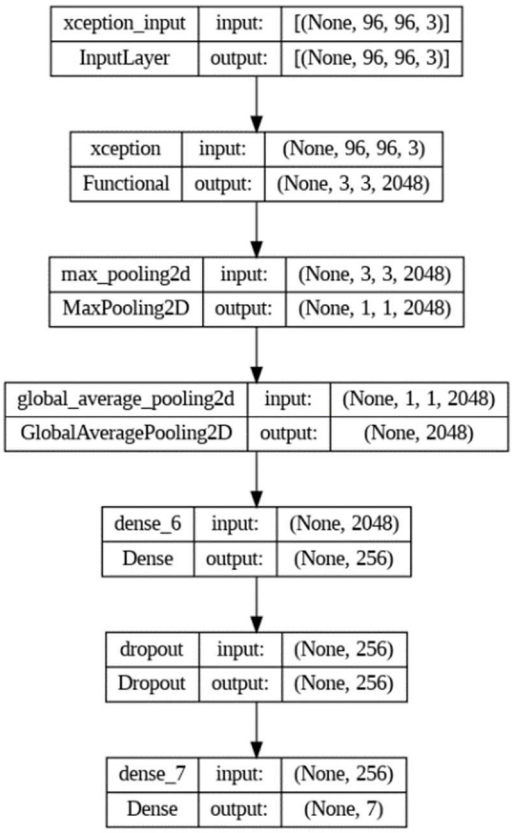

Fig. 12 Xception architecture modified for FER

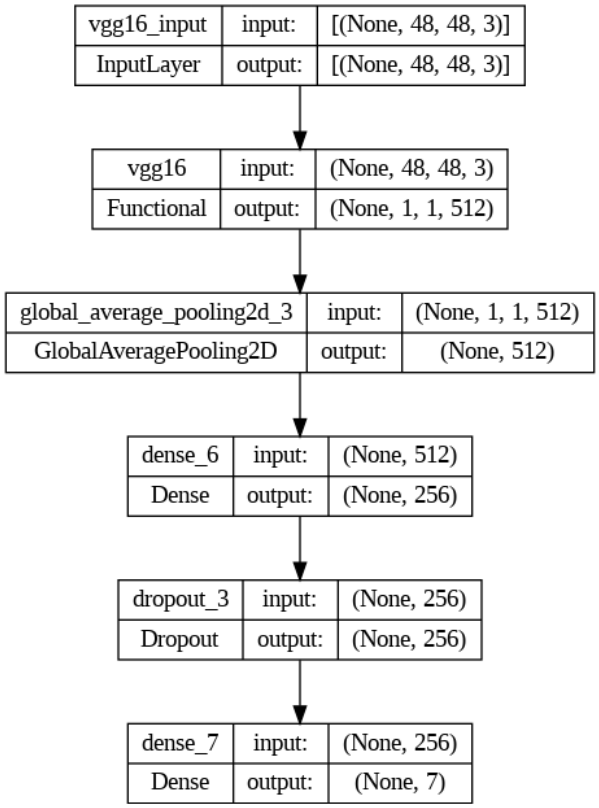

Fig. 13 VGG16 architecture modified for FER

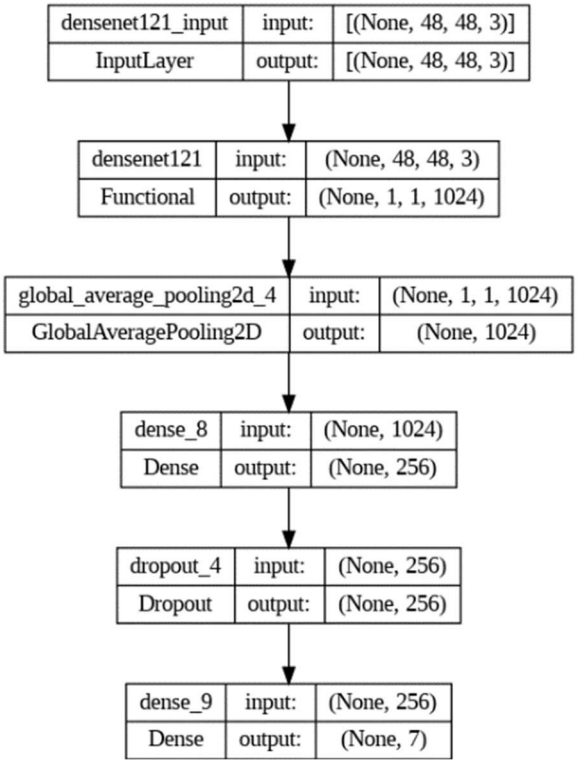

Fig. 14 Densenet121 architecture modified for FER

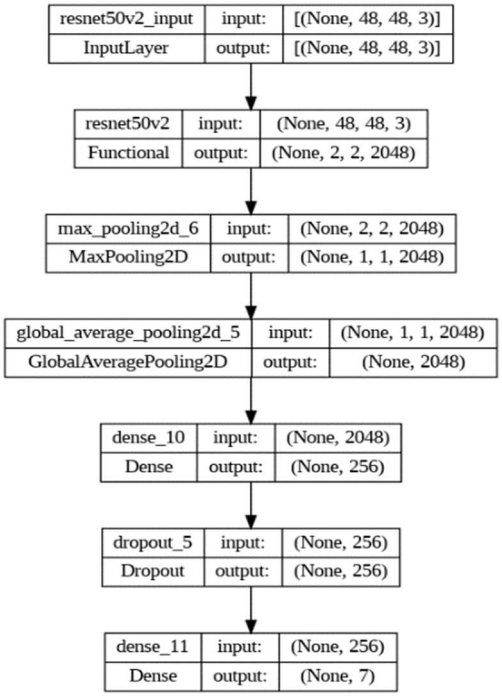

Fig. 15 Resnet50v2 architecture modified for FER

Dense is a fully connected layer. This last fully connected layer can accept the previously extracted features and output the probability for each class. Afterward, neurons are removed using Dropout to prevent the network from overfitting. Finally, it was compressed again into 7 outputs, designed to avoid losing as many of each model's characteristics as possible. The detailed layers for each model are shown in the following pictures.

D. Materials

The hardware and software used for the study are as follows. CPU Intel(R) Core (TM) i7-12700KF, GPU GTX 1060 3GB / Windows 10, python 3.8.10, tensorflow-gpu 2.10.0, jupyterlab 2.2.6, numpy 1.19.2, pandas 1.1.3, pillow 8.0 .1, matplotlib 3.3.2, sklearn 1.3.2, NVIDA cuda 11.2, NVIDA cudnn 8.1

CNN model architecture Each input image consists of 3 gray RGB channels with a size of 48*48. Before training, pre-trained 'ImageNet' weights are used. To respond to various image inputs, the image generator prepares for various input sizes of faces at a magnification of 0.3, angle 20, tilt 0.3, and movement 0.2. This option applies equally to all image generators.

Each model was trained with batch sizes of 64 and 10 and epoch 30. Learning was conducted by giving weights to each class according to the data imbalance in the dataset. The class weight formula is as follows.

$$Class\ weight = \{class\ id : round(\frac{\max(Counter(C).values())}{N}, 2|for(class\ id, N) \in Counter(C).items()\}$$

The formula for evaluating each model is as follows.

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

$$Recall = \frac{(TP)}{(TP + FN)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

$$F1\ Score = \frac{2 * (recall * Precision)}{Recall * Precision}$$

TP = True-Positive, TN = True-Negative, FP = False-Positive and FN = False-Negative

III. Results and Discussions

A. Evaluation and Result

Parameters for each model are as shown in Table 2. The prediction rates of the models are shown in Table 3 below. Each model has an output convolutional network with the same structure. Additionally, the dataset being learned is also the same. The code and results of the above experiments can be found at the following address[1]. The analysis of these results is as follows. In ImageNet, models with higher prediction rates also had relatively higher FER classification.

TABLE II
CNN models Parameters values

No	Parameters			
	Model	**Total**	**Trainable**	**None**
1	MobileNet	02.6M	02.5M	00.1M
2	Xception	21.3M	21.3M	00.5M
3	VGG16	14.8M	14.8M	0
4	DenseNet121	07.3M	07.2M	00.1M
5	ResNet50V2	24.1M	24.0M	00.1M

[1] https://github.com/ljh77/korean_FER_24

The model that showed the highest KFE prediction rate was the model learned based on Xception (2-3) from FER2013, CK+, and KFE. In other words, in the learning of artificial intelligence image models, it can be seen that the prediction rate is affected by the inclusion of many and diverse datasets and the actual prediction images that are subject to classification.

TABLE III
ACCURACY THAT EACH CNN MODELS

No	Model	Dataset	FER2013Acc	KFE Acc	JAFFE Acc
1-1	MobileNet	FER2013	0.371	0.175	0.225
1-2	MobileNet	FER2013, CK+	0.354	0.137	0.235
1-3	MobileNet	FER2013, CK+,KFE	0.352	0.190	0.258
1-4	MobileNet	CK+*	0.135	0.146	0.169
1-5	MobileNet	KFE*	0.157	0.144	0.160
2-1	Xception	FER2013	0.604	0.264	0.272
2-2	Xception	FER2013, CK+	0.600	0.270	0.263
2-3	Xception	FER2013, CK+,KFE	0.619	**0.328**	**0.315**
2-4	Xception	CK+*	0.298	0.177	0.235
2-5	Xception	KFE*	0.326	0.240	0.192
3-1	VGG16	FER2013	**0.627**	0.235	0.291
3-2	VGG16	FER2013, CK+	0.621	0.236	0.277
3-3	VGG16	FER2013, CK+,KFE	**0.626**	0.314	**0.324**
3-4	VGG16	CK+*	0.285	0.158	0.254
3-5	VGG16	KFE*	0.371	**0.305**	0.286
4-1	DenseNet121	FER2013	0.497	0.252	0.300
4-2	DenseNet121	FER2013, CK+	0.509	0.257	0.300
4-3	DenseNet121	FER2013, CK+,KFE	0.529	0.253	0.376
4-4	DenseNet121	CK+*	0.169	0.153	0.174
4-5	DenseNet121	KFE*	0.178	0.150	0.174
5-1	ResNet50V2	FER2013	0.492	0.194	0.225
5-2	ResNet50V2	FER2013, CK+	0.491	0.194	0.224
5-3	ResNet50V2	FER2013, CK+,KFE	0.507	0.194	0.192
5-4	ResNet50V2	CK+*	0.156	0.166	0.155
5-5	ResNet50V2	KFE*	0.145	0.132	0.131

*BATCH SIZE 10

Overall, the more diverse data you learn, the higher the prediction rate. This is the same as a typical CNN learning prediction. Even though MobileNet has about 1/10 the number of parameters compared to other models, its performance was about 50% of that of other models. This also shows similarities to other image studies and comparative studies between models. VGG16 and Xception gave similar prediction rates. Xception has about twice as many parameters as VGG16, and this result can be seen because the classes of images that need to be distinguished are not diverse, so the intensive neural network has less influence. It can be predicted that the less diversity of images and the fewer classes to be classified, the more advantageous a lightweight neural network is over a very deep neural network. This is the same as the results of other studies.

IV. CONCLUSION

Through the study, among the five CNN models, the VGG16 model had the highest prediction rate on the FER2013 confirmation data, followed closely by the Xception model. However, there was no significant difference in prediction rates between the two models. Resnet and Desnet were next, but there was also no significant difference. The high prediction rates of KFE and JAFFE, which are composed of East Asians, are the 2-3 and 3-3 models learned on all datasets, as well as the Xception and VGG16 models. VGG16 showed a high prediction rate despite being a primitive model compared to other models. This can be inferred that the human facial emotion classification image dataset is less complex than other image datasets. Complex layered models such as DenseNet or ResNet showed higher performance than Moblienet.

In other words, it can be seen that Xception, which shows high performance in ImageNet classification and FER classification, can be used for image classification. This also showed the same research trend in other image classification studies. VGG16 can be used for classification on images with few classes and little change. It was confirmed again that the dataset is diverse, and that the prediction rate increases as the number increases. However, even when learning with FER2013, the prediction rate for the KFE and JAFFE datasets was not high. This appears to be because there is a lot of image data in FER2013. In the case of Trained 2-5 and 3-5 trained with KFE, all three datasets showed similar prediction rates. In other words, the learned weight converges to the average value of the learned data, so if the learning image is biased, the prediction rate is low for images that do not correspond to the data. As a result, for accurate classification of images, they must be learned from uniform and diverse datasets.

In this way, we compared the image classification prediction rates according to datasets and models and explored their significance. This study will be helpful in constructing datasets and setting up models in the field of image classification and computer vision in the future. As future research, we will study lightweight image classification models for specific tasks and classifications and continue research on uniform composition of datasets.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1I1A305223413).

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[3] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/taffc.2020.2981446.

[4] Y. Li, D. Niu, and J. Peng, "Boundary-Aware Face Alignment with Enhanced HourglassNet and Transformer," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023, doi:10.1561/116.00000115.

[5] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP," Sensors, vol. 23, no. 9, p. 4204, Apr. 2023, doi: 10.3390/s23094204.

[6] M. Wu, J. Zhou, Y. Peng, S. Wang, and Y. Zhang, "Deep Learning for Image Classification: A Review," Proceedings of 2023 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2023), pp. 352–362, 2024, doi: 10.1007/978-981-97-1335-6_31.

[7] S. M. Saleem Abdullah and A. M. . Abdulazeez, "Facial Expression Recognition Based on Deep Learning Convolution Neural Network: A Review", jscdm, vol. 2, no. 1, pp. 53–65, Apr. 2021, Accessed: Jun. 03, 2024. [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/7906.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, doi:10.1109/cvprw.2010.5543262.

[9] Korea Institute for Artificial Intelligence & Society (KIAI). "AIhub." [Online] Available: https://aihub.or.kr/.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, doi: 10.1109/cvpr.2001.990517.

[11] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, doi: 10.1109/afgr.1998.670949.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/tpami.2017.2699184.

[13] Tan, Mingxing; Le, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International conference on machine learning. PMLR*, 2019. p. 6105-6114.

[14] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, doi: 10.1109/cvpr.2019.00065.

[15] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning Augmentation Strategies From Data," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, doi: 10.1109/cvpr.2019.00020.

[16] A. Howard et al., "Searching for MobileNetV3," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, doi: 10.1109/iccv.2019.00140.

[17] Tan, Mingxing; Le, Quoc V. Mixconv: Mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595, 2019.

[18] Bochkovskiy, Alexey; Wang, Chien-Yao; Liao, Hong-Yuan Mark. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[19] K. Han et al., "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/tpami.2022.3152247.

[20] Dai, Zihang, et al. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 2021, 34: 3965-3977.

[21] Chen, Xi, et al. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794, 2022.

[22] Wortsman, Mitchell, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. PMLR, 2022. p. 23965-23998.

[23] S. Srivastava and G. Sharma, "OmniVec: Learning robust representations with cross modal sharing," *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2024, doi: 10.1109/wacv57701.2024.00127.

[24] M. Baharani, S. Mohan, and H. Tabkhi, "Real-Time Person Re-identification at the Edge: A Mixed Precision Approach," *Image Analysis and Recognition*, pp. 27–39, 2019, doi: 10.1007/978-3-030-27272-2_3.

[25] Y. Duan, J. Lu, and J. Zhou, "UniformFace: Learning Deep Equidistributed Representation for Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, doi: 10.1109/cvpr.2019.00353.

[26] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards Universal Representation Learning for Deep Face Recognition," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, doi: 10.1109/cvpr42600.2020.00685.

[27] R. Adityatama and A. T. Putra, "Image classification of Human Face Shapes Using Convolutional Neural Network Xception Architecture with Transfer Learning," Recursive Journal of Informatics, vol. 1, no. 2, pp. 102–109, Sep. 2023, doi: 10.15294/rji.v1i2.70774.

[28] Z. Akhtar, M. R. Mouree, and D. Dasgupta, "Utility of Deep Learning Features for Facial Attributes Manipulation Detection," *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)*, Sep. 2020, doi: 10.1109/hccai49649.2020.00015.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, doi: 10.1109/cvpr.2009.5206848.

[30] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Apr. 2015, doi: 10.1007/s11263-015-0816-y.