Knowledge Engineering and Data Science (KEDS)
Vol 5, No 1, December 2022, pp. 78–86

pISSN 2597-4602
eISSN 2597-4637

Human Facial Expressions Identification using Convolutional Neural Network with VGG16 Architecture

Luther Alexander Latumakulita [a,1,*], Sandy Laurentius Lumintang [a,2], Deiby Tineke Salaki [a,3], Steven R. Sentinuwo [a,4], Alwin Melkie Sambul [a,5], Noorul Islam [b,6]

[a] *Sam RatulangiUniversity, Jalan Kampus, Manado 95115, Indonesia*
[b] *Kanpur Institute of Technology, A-1, UPSIDC, Rooma Industrial Area, Kanpur, 201008, India*
[1] *latumakulitala@unsrat.ac.id;* [2] *sandy.laurentius@gmail.com;* [3] *deibyts.mat@unsrat.ac.id,* [4] *steven@unsrat.ac.id,* [5] *asambul@unsrat.ac.id,* [6] *noorul.islam3101@gmail.com*
* *latumakulitala@unsrat.ac.id*

ARTICLE INFO

Article history:
Received 30 May 2022
Revised 30 June 2022
Accepted 14 October 2022
Published online 7 November 2022

Keywords:
CNN
Deep Learning
Facial Expressions Identification
VGG16

ABSTRACT

The human facial expression identification system is essential in developing human interaction and technology. The development of Artificial Intelligence for monitoring human emotions can be helpful in the workplace. Commonly, there are six basic human expressions, namely anger, disgust, fear, happiness, sadness, and surprise, that the system can identify. This study aims to create a facial expression identification system based on basic human expressions using the Convolutional Neural Network (CNN) with a 16-layer VGG architecture. Two thousand one hundred thirty-seven facial expression images were selected from the FER2013, JAFFE, and MUG datasets. By implementing image augmentation and setting up the network parameters to Epoch of 100, the learning rate of 0,0001, and applying in the 5Fold Cross Validation, this system shows performance with an average accuracy of 84%. Results show that the model is suitable for identifying the basic facial expressions of humans.

I. Introduction

Humans can produce different facial expressions [1], but some distinctive facial configurations are associated with specific emotions [2], regardless of gender [3], age [4], cultural background [5], and socialization history [6]. Facial expressions accounted for 55% of message delivery, while language and voice accounted for 7% and 38%, respectively [7]. Universally, six basic expressions have been put forward in Ekman and Friesen's research, namely anger, disgust, happiness, sadness, and surprise expressions.

Along with the developments of technology, the interaction between humans and technology plays a vital role in daily activities. Artificial intelligence can facilitate work and help humans make decisions based on the results of their analysis. One example of the application of this technology is to identify human facial expressions. Some services currently use scoring systems by manually selecting on a computer display, but these systems are considered inappropriate for showing expressions of customer satisfaction [8]. In addition, facial expression identification systems can be developed and applied in various fields, such as psychological patient emotion detection, lie detection, security system with face recognition, entertainment recommendations according to emotions (movies, music, tourist attractions, shopping products), robot development, monitoring system of an employee's facial expressions when interacting with Customers and so on.

The development of facial expression identification technology can use the deep learning method that makes a computer learn from the depths of an image and identify it. One type of deep learning method currently the most significant in image recognition is the Convolutional Neural Network (CNN) with a 16-layer Visual Geometry Group (VGG) architecture. Sang et al. [9] showed an average test accuracy of 71.9% using CNN's deep BKVGG12. Gultom et al. [10] showed an average accuracy

https://doi.org/10.17977/um018v5i12022p78-86

of 89±7% using VGG16 transfer learning for batik classification. Porcu et al. [11] found that applying image augmentation can significantly improve test acumen compared to previous studies. Caroppo et al. [12] found that using CNN's deep learning VGG16 architecture for facial expression identification showed the highest accuracy compared to other architectures.

CNN is one part of deep feedforward artificial neural networks (ANN) widely applied to computer vision, also known as ConvNet, and has an architecture derived from nodes or neurons connected at a layer [13]. In general, the types of layers on CNN are divided into the feature extraction layer and the classification layer.

In this study, the CNN-based Deep Learning method will be used to identify six basic expressions of the human face with the VGG16 architecture to get good enough accuracy. Image augmentation will be applied to the image data, and then the data will be trained using the K-Fold Cross Validation method, which produces a confusion matrix due to its evaluation. In the future, the intelligent model proposed in this research can be implemented in a control system that needs human expression, like an automated gate or system surveillance.

II. Methods

In this study, the identification of basic expressions of the human face consisted of several stages, as seen in Figure 1.

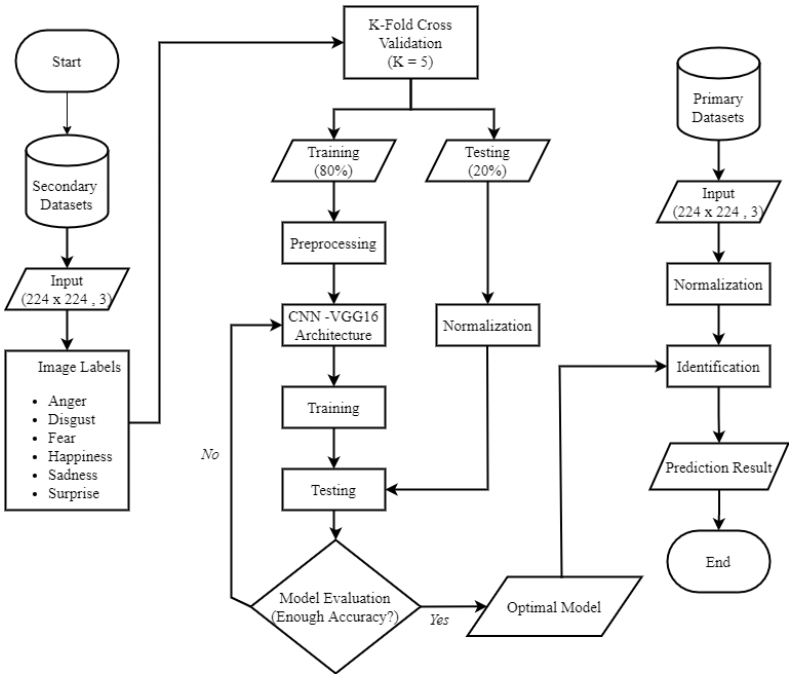

Fig. 1. Research flowchart

The selecting dataset stage is the process of collecting the dataset. The secondary data collected is determined one by one by looking at the accuracy of expressions based on physical descriptions of basic human facial expressions and image clarity (no watermarks or other objects hinder facial clarity). Human facial expression image data has been selected from open-source datasets (FER2013, JAFFE, and MUG).

Facial expressions to be trained and identified are the six basic human expressions based on physical descriptions or criteria as follows [14]. Anger; Brows wrinkled, eyes wide, lips tightened and pressed together. Disgust; Eyebrows fall, eyes narrow, nose wrinkles, lips split, jaw drop. Fear; Eyebrows raised and pulled together, upper eyelids raised, lower eyelids tense, lips parted and stretched. Happiness; Eyes narrowed and wrinkles around him, cheeks raised, lips pulled back, showing teeth in a smile. Sadness; Eyebrows are knitted, eyes are slightly closed, the corners of the

lips are depressed, and the lower lip is raised. Surprise; Eyebrows raised, upper eyelids raised, lips parted, jaws dropped. The result for each label is shown in Table 1.

Table 1. Image data of facial expressions

No	Facial Expressions	Data
1	Anger	325
2	Disgust	216
3	Fear	197
4	Happiness	758
5	Sadness	260
6	Surprise	381

In Table 1, the human facial expression image data that has been selected amounts to 2137 images, where anger expressions amounted to 325 data, expressions of disgust amounted to 216 data, expressions of fear amounted to 197 data, expressions of happiness amounted to 758 data, sadness expressions amounted to 260 data, and expressions of surprise amounted to 381 data.

This data input stage retrieves data from directories that are then labeled accordingly for each image, and the data is inputted using sizes 224 × 224 and channel 3 (RGB). Before entering the training stage, the data will be divided first using K-fold Cross Validation with K = 5, resulting in training and testing data with a ratio of 80:20, which means the model trains the training dataset 5 times with different training data for each fold.

The training data entered will be normalized, and applied image augmentation for each data at the preprocessing stage. Data normalization is a linear scale technique for changing the pixel scale of an image from 0 to 1. The entered image data will be divided by 255 (RGB range 0-255). Paper [15] suggests that image augmentation increases the size and diversity of existing training pools without manually collecting new data. This process generates additional training data from existing examples by adding them using random transformations that produce impressions that appear trustworthy. Image augmentation is helpful so that computers can learn more about the data that has been trained from various points of view and multiply the data to be prepared. Image augmentation is applied by performing a series of random preprocessing transformations to existing data, such as flipping horizontally and vertically, tilting, cropping, cropping, zooming in and out, and rotating. The image augmentation applied is shown in Table 2.

Table 2. The image augmentation applied

Parameter	Value	Detail
Rotation range	40	the image is rotated at an angle of 0.2 degrees
Width shift range	0.25	the image's width is shifted with an angle of 0.25 degrees
Height shift range	0.25	the height of the image is shifted at an angle of 0.25 degrees
Shear range	0.20	the image is shifted clockwise by 0.2 degrees
Zoom range	0.2	the image is enlarged by 1 + 0.2 from the area of the image
Horizontal flip	True	the image is rotated horizontally
Fill mode	nearest	the image pixels lost during changes will be filled with the nearest pixel value to maintain the integrity of the image quality

The training dataset preprocessed before will be entered into the VGG16 architecture model and evaluated using the testing dataset divided before for each fold. VGG 16 is one of the CNN architectures, which was put forward by Simonyan and Zisserman when competing in the ImageNet Large Scale Visual Recognition Challenge and making the top-5 with an accuracy of 92.7% [16]. VGG16 is an improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolution layers, respectively) with multiple 3x3 kernel-sized filters one after another [17]. The VGG16 architecture can be seen in Figure 2.

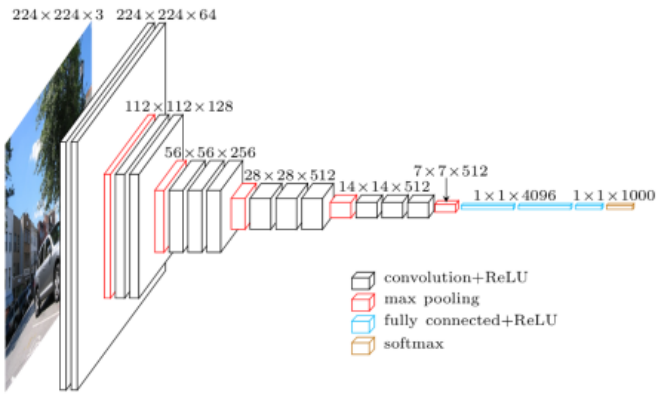

Fig. 2. VGG16 architecture

Figure 2 shows 16 layers with 13 convolution layers and 3 Fully Connected layers. In 13 convolution layers are given a filter of 64 for the first two layers, a filter of 128 for the following two layers, a filter of 256 for the subsequent three layers, a filter of 512 for the following three layers, and a filter of 512 for the subsequent three layers with the max-pooling layer in each filter change, which amounts to 5 layers. The image is inputted using a size of 224 x 224 x 3, which indicates that the model will read the image with a size of 224 x 224 with channel 3, namely RGB (Red, Green, Blue).

MaxPooling layer uses a size of 2 x 2 and stride two so that it changes the image size, which was originally 224 x 224, and produces a feature map of 112 x 112 (filter 128), then 56 x 56 (filter 256), then to 28 x 28 (filter 512), then to 14 x 14 (filter 512) and finally to 7 x 7. The pooling layer is a way to reduce matrix size to speed up computing and easily control overfitting. One way to use this pooling layer is to apply MaxPooling. MaxPooling is a function that selects the maximum value of a window region and is then represented as a new pixel [18].

Use two dense layers of 4096 (with two dropouts (0.5) layers) and one softmax layer for three fully connected layers. Dropout temporarily eliminates a neuron in the network's Hidden Layer or Visible Layer [19]. The model obtained will then be assessed to determine whether the accuracy is enough. If the accuracy is still lacking, changes will be made to the parameters of the epoch network, learning rate, and batch size until the accuracy obtained is satisfactory.

At the identification stage, the primary data obtained will be tested using the model with the best accuracy. The entered data will be normalized and resized to 224x224 with channel 3 (RGB). The last stage is the prediction of the result of the identification stage and is evaluated using the confusion matrix. A confusion matrix is a table frequently used to evaluate the effectiveness of a classification model [20]. In order to assess the accuracy of a model's predictions, the confusion matrix compares the predicted labels against the actual labels.

We may construct various model performance measures with these four results, including accuracy, precision, recall, and F1-score [21]. These measures give a more detailed picture of the model's performance than the model's overall accuracy rate alone. Overall, a confusion matrix is a valuable tool for evaluating the performance of a classification model and finding potential areas for improvement.

III. Results and Discussion

Testing is done by applying 5-Fold Cross Validation, where the data is divided into five different sets, which are then carried out the testing process five times. The testing process is done by setting the epoch value = 100, learning rate = 0.0001, and batch size = 32. Each fold of the model obtained performed performance testing by applying a confusion matrix to data testing. Here are the results of testing accuracy in each fold.

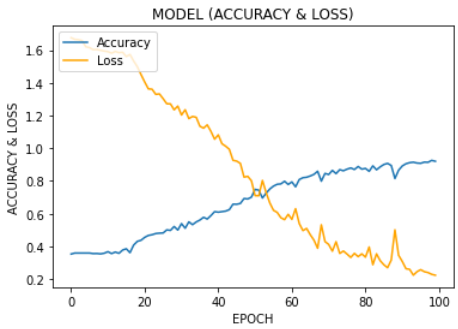

Fig. 3. Plot training fold 1

Figure 3 shows the results of plot accuracy and loss on fold 1 with a total of 1794 data trains, where the accuracy is increasing close to the number 1.0, with the highest value reaching 92%. The loss is getting closer to 0, with the lowest value of 0.224.

Table 3. Confusion matrix fold 1

Fold 1	Anger	Disgust	Far	Happiness	Sadness	Surprise
Anger	**49**	0	0	0	4	0
Disgust	3	**31**	0	3	0	0
Fear	2	0	**17**	2	5	1
Happiness	0	1	1	**111**	0	0
Sadness	9	0	0	4	**35**	0
Surprise	0	0	5	2	0	**58**
Accuracy				87.7%		
Precision				88%		
F1 Score				87.6%		

Based on Table 3, the red labeled numbers are the data predicted to be correct based on the test results of the fold one training model. The confusion matrix from a total of 343 data testing has an accuracy of 87.7%, a precision of 88%, and an F1 Score of 87.6%.

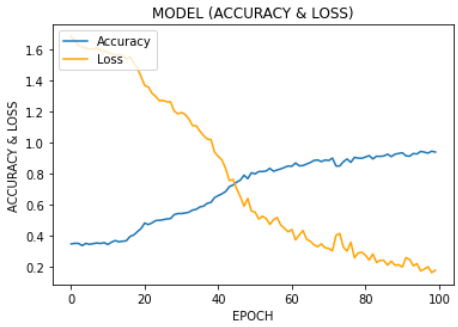

Fig 4. Plot training fold 2

Figure 4 shows the results of plot accuracy and loss on fold 2 with a total of 1795 data trains, where the accuracy is increasing close to the number 1.0, with the highest value reaching 94%. The loss is getting closer to 0, with the lowest value of 0.160.

Table 4. Confusion matrix fold 2

Fold 2	Anger	Disgust	Far	Happiness	Sadness	Surprise
Anger	**42**	0	2	0	2	0
Disgust	0	**33**	0	1	2	0
Fear	0	3	**25**	1	6	2
Happiness	0	0	0	**133**	0	0
Sadness	2	0	0	0	**33**	0
Surprise	1	0	5	0	0	**49**
Accuracy				92.1%		
Precision				92.2%		
F1 Score				92.0%		

Based on Table 4, the red labeled numbers are the data predicted to be correct based on the test results of the fold-two training model. The confusion matrix from a total of 342 data testing has an accuracy of 92.1%, a precision of 92.2%, and an F1 Score of 92.0%.

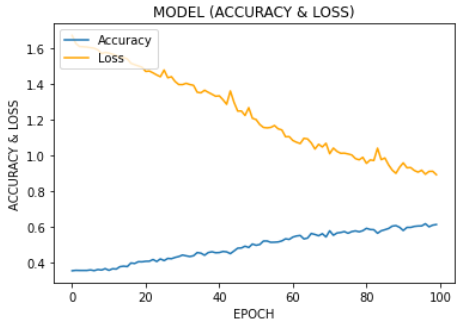

Fig 5. Plot training fold 3

Figure 5 shows the results of plot accuracy and loss on fold 3 with a total of 1795 data trains, where the accuracy is increasing close to the number 1.0, with the highest value reaching 61%. The loss is getting closer to the number 0, with the lowest value of 0.892.

Table 5. Confusion matrix fold 3

Fold 3	Anger	Disgust	Far	Happiness	Sadness	Surprise
Anger	**25**	1	0	26	2	0
Disgust	6	**7**	1	19	0	0
Fear	3	1	**14**	18	0	1
Happiness	4	0	0	**116**	0	0
Sadness	5	0	2	22	**13**	0
Surprise	5	0	0	22	0	**29**
Accuracy				59.6%		
Precision				69.3%		
F1 Score				56.9%		

Based on Table 5, the red labeled numbers are the data predicted to be correct based on the test results of the fold three training model. The confusion matrix from a total of 342 data testing has an accuracy of 59.6%, a precision of 69.3%, and an F1 Score of 56.9%.

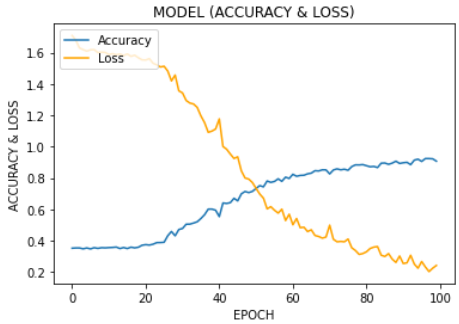

Fig. 6. Plot training fold 4

Figure 6 shows the results of plot accuracy and loss on fold 4 with a total of 1795 data trains, where the accuracy is increasing close to the number 1.0, with the highest value reaching 92%. The loss is getting closer to 0, with the lowest value of 0.202.

Table 6. Confusion matrix fold 4

Fold 4	Anger	Disgust	Far	Happiness	Sadness	Surprise
Anger	**50**	0	2	0	8	0
Disgust	1	**36**	1	0	2	0
Fear	0	1	**20**	2	0	0
Happiness	1	2	2	**118**	1	0
Sadness	3	0	0	1	**30**	0
Surprise	0	0	4	1	0	**56**
Accuracy				90.6%		
Precision				91.6%		
F1 Score				90.9%		

Based on Table 6, the red labeled numbers are the data predicted to be correct based on the test results of the fold-four training model. The confusion matrix from a total of 342 data testing has an accuracy of 90.6%, a precision of 91.6%, and an F1 Score of 90.9%.

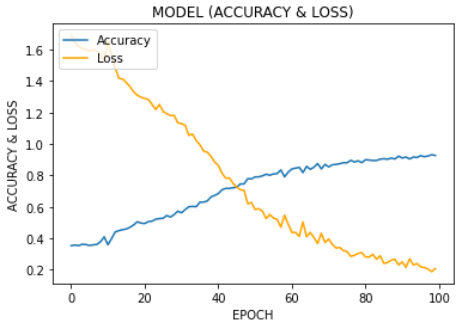

Fig 7. Plot training fold 5

Figure 7 shows the results of plot accuracy and loss on fold 5 with a total of 1795 data trains, where the accuracy is increasing close to the number 1.0, with the highest value reaching 93%. As for the loss is getting closer to the number 0, with the lowest value being 0.187.

Table 7. Confusion matrix fold 5

Fold 5	Anger	Disgust	Far	Happiness	Sadness	Surprise
Anger	**42**	1	0	0	4	0
Disgust	1	**24**	0	0	2	0
Fear	0	0	**25**	1	5	3
Happiness	0	2	0	**113**	1	1
Sadness	2	0	0	1	**46**	0
Surprise	0	0	3	0	0	**65**
Accuracy				92.1%		
Precision				92.4%		
F1 Score				92.1%		

Based on Table 7, the red labeled numbers are the data predicted to be correct based on the test results of the fold-five training model. The confusion matrix from a total of 342 data testing has an accuracy of 92.1%, a precision of 92.4%, and an F1 Score of 92.1%. The average accuracy of each fold can be seen in Table 8.

Table 8. Average training accuracy

Fold-K	1	2	3	4	5
Accuracy	87.7%	92.1%	59.6%	90.6%	92.1%
Average			84.4%		

From Table 8, fold accuracy is obtained from 1 to 5 with an average of 84.4%, where the highest accuracy is received in the second and fifth folds at 92.1%, and the lowest accuracy is 59.6% in the third fold. The best model that will be used for the identification stage is the fold 2 model. Figure 8 shows the plotting graphic of accuracies between training and testing processes for all folds.

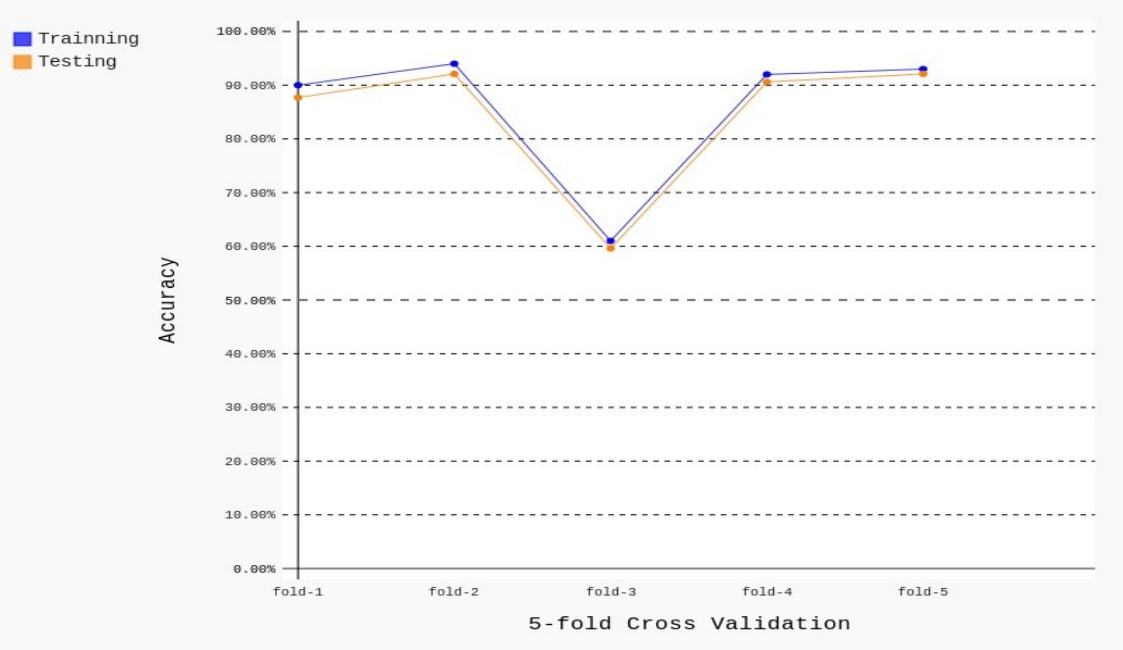

Fig 8. Accuracies comparison between training and testing processes

Table 9. Identification results

Indicator	Facial Expressions					
	Anger	***Disgust***	***Far***	***Happiness***	***Sadness***	***Surprise***
True	6	5	4	6	6	4
False	0	1	2	0	0	2
Accuracy			**86.1%**			

From Table 9, identification of primary data was carried out and found from 36 data, 31 of which were predicted to be correct and five were predicted to be correct. The accuracy of the confusion matrix is 86.1%.

IV. Conclusion

CNN with the VGG16 architecture model managed to identify the primary expressions of the human face with an epoch count of 100, a learning rate of 0.0001, and a batch size of 32, resulting in an average accuracy of the 1st to fifth fold is 84.4% with an average test data accuracy of 86.1%, so it can be said that the model built is stable and good enough to use. In the future, we will develop a system to control automatic gates at Sam Ratulangi University. The gate will automatically open after receiving the best smile from people who want to enter the University area.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] D. L. Z. Astuti, S. Samsuryadi, and D. P. Rini, "Real-Time Classification Of Facial Expressions Using A Principal Component Analysis And Convolutional Neural Network," SINERGI, vol. 23, no. 3, p. 239, Oct. 2019.

[2] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," Psychol. Sci. Public Interes., vol. 20, no. 1, pp. 1–68, Jul. 2019.

[3] Y. Park and M. Garcia, "Pedestrian safety perception and urban street settings," Int. J. Sustain. Transp., vol. 14, no. 11, pp. 860–871, Sep. 2020.

[4] [4] S. Simpson, L. Richardson, G. Pietrabissa, G. Castelnuovo, and C. Reid, "Videotherapy and therapeutic alliance in the age of COVID‐19," Clin. Psychol. Psychother., vol. 28, no. 2, pp. 409‐421, Mar. 2021.

[5] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures.," Emotion, vol. 18, no. 1, pp. 75–93, Feb. 2018.

[6] A. J. Umaña‐Taylor and N. E. Hill, "Ethnic‐Racial Socialization in the Family: A Decade's Advance on Precursors and Outcomes," J. Marriage Fam., vol. 82, no. 1, pp. 244‐271, Feb. 2020.

[7] S. M. Saleem Abdullah and A. M. Abdulazeez, "Facial Expression Recognition Based on Deep Learning Convolution Neural Network: A Review," J. Soft Comput. Data Min., vol. 02, no. 01, Apr. 2021.

[8] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of Micro-Expressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi," in 2020 Fifth International Conference on Informatics and Computing (ICIC), Nov. 2020, pp. 1–9.

[9] D. V. Sang, N. Van Dat, and D. P. Thuan, "Facial expression recognition using deep convolutional neural networks," in 2017 9th International Conference on Knowledge and Systems Engineering (KSE), Oct. 2017, pp. 130–135.

[10] Y. Gultom, A. M. Arymurthy, and R. J. Masikome, "Batik Classification using Deep Convolutional Network Transfer Learning," J. Ilmu Komput. dan Inf., vol. 11, no. 2, p. 59, Jun. 2018.

[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, Nov. 2020.

[12] A. Caroppo, A. Leone, and P. Siciliano, "Comparison Between Deep Learning Models and Traditional Machine Learning Approaches for Facial Expression Recognition in Ageing Adults," J. Comput. Sci. Technol., vol. 35, no. 5, pp. 1127–1146, Oct. 2020.

[13] C. Modarres, N. Astorga, E. L. Droguett, and V. Meruane, "Convolutional neural networks for automated damage recognition and damage type identification," Struct. Control Heal. Monit., vol. 25, no. 10, p. e2230, Oct. 2018.

[14] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional Expression: Advances in Basic Emotion Theory," J. Nonverbal Behav., vol. 43, no. 2, pp. 133–160, Jun. 2019.

[15] G. Ramirez-Gargallo, M. Garcia-Gasulla, and F. Mantovani, "TensorFlow on State-of-the-Art HPC Clusters: A Machine Learning use Case," in 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), May 2019, pp. 526–533.

[16] S. A. Asmai*, M. N. D. Mohamad Zukhairin, A. S. M. Jaya, A. F. N. Abdul Rahman, and Z. B. Abal Abas, "Mosquito Larvae Detection using Deep Learning," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 804–809, Oct. 2019.

[17] R. Jain, P. Nagrath, G. Kataria, V. Sirish Kaushik, and D. Jude Hemanth, "Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning," Measurement, vol. 165, p. 108046, Dec. 2020.

[18] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types," Comput. Civ. Infrastruct. Eng., vol. 33, no. 9, pp. 731–747, Sep. 2018.

[19] M. Elleuch, R. Maalej, and M. Kherallah, "A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition," Procedia Comput. Sci., vol. 80, pp. 1712–1723, 2016.

[20] A. P. Wibawa, S. A. Kurniawan, and I. A. E. Zaeni, "Determining Journal Rank by Applying Particle Swarm Optimization-Naive Bayes Classifier," J. Inf. Technol. Manag., vol. 13, no. 4, 2021.

[21] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons & Fractals, vol. 140, p. 110120, Nov. 2020.