Jurnal Elektro
p-ISSN: 1979-9780
e-ISSN: 2746-4288

Jurnal Elektro Vol. 18 No. 1 April 2025
journal homepage: https://ejournal.atmajaya.ac.id/index.php/JTE

# Classification Of Multi-Class Face Expression Using Modification Of VGG-16 Model

**Aryadana Priyatama[1], Sugiyanto Sugiyanto[2]**

[1]Informatics Engineering Study Program, Dian Nuswantoro University
[2]Distance Learning Informatics Study Program, Dian Nuswantoro University

Article Info

Article history:

Received
26-05-2025

Accepted
04-06-2025

Keywords:
Facial recognition, Facial expression classification, VGG-16, GridSearch, K-Fold Cross Validation

## Abstract

In the era of modern technology, facial recognition has become an important application in various fields, such as security, education and health. One method used to recognize faces is a Convolutional Neural Network (CNN), specifically the VGG-16 architecture which is known for its consistent performance. But even though CNN can recognize faces, its accuracy in recognizing faces is inadequate. This research aims to increase the accuracy of facial expression classification so that it is more optimal by modifying the CNN VGG-16 architecture. This research uses GridSearch techniques, K-Fold Cross Validation, and utilizes multiple datasets. The dataset used consists of two image datasets, namely SMIC and SAMM facial-micro expressions, each of which has been normalized and converted to a grayscale scale measuring 48x48 pixels. The GridSearch process is applied to optimize parameters such as the number of filters, learning rate, dropout rate, activation function, and batch size. The K-Fold Cross Validation technique with five folds was used to ensure the generalization of the model to new data. The research results show that this modification is able to achieve validation accuracy of up to 98.31% in the training process, showing a significant improvement compared to the standard method. And showed an increase in accuracy in testing of 98.04% in research.

Artikel Info

Histori Artikel:

Diterima
26-05-2025

Disetujui
04-06-2025

Kata Kunci:
GridSearch, Klasifikasi ekspresi wajah, K-Fold Cross Validation, Pengenalan wajah, VGG-16

## Abstrak

Di era teknologi modern, pengenalan wajah telah menjadi aplikasi penting di berbagai bidang, seperti keamanan, pendidikan, dan kesehatan. Salah satu metode yang digunakan untuk mengenali wajah adalah Convolutional Neural Network (CNN), khususnya arsitektur VGG-16 yang dikenal dengan kinerjanya yang konsisten. Namun meskipun CNN dapat mengenali wajah, akurasinya dalam mengenali wajah belum memadai. Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi ekspresi wajah agar lebih optimal dengan memodifikasi arsitektur CNN VGG-16. Penelitian ini menggunakan teknik GridSearch, K-Fold Cross Validation, dan memanfaatkan beberapa dataset. Dataset yang digunakan terdiri dari dua dataset citra, yaitu ekspresi wajah-mikro SMIC dan SAMM yang masing-masing telah dinormalisasi dan dikonversi ke skala grayscale berukuran 48x48 piksel. Proses GridSearch diterapkan untuk mengoptimalkan parameter seperti jumlah filter, learning rate, dropout rate, fungsi aktivasi, dan ukuran batch. Teknik K-Fold Cross Validation dengan lima kali lipat digunakan untuk memastikan generalisasi model ke data baru. Hasil penelitian menunjukkan bahwa modifikasi ini mampu mencapai akurasi validasi hingga 98,31% pada proses pelatihan, menunjukkan peningkatan yang signifikan dibandingkan dengan metode standar. Dan menunjukkan peningkatan akurasi pada pengujian sebesar 98,04% pada penelitian.

*Corresponding author. Sugiyanto
Email address: sugiyanto@dsn.dinus.ac.id

## 1. INTRODUCTION

Facial expressions play a crucial role in daily human communication as a non-verbal channel for conveying emotions. Through facial cues, individuals can express internal emotional states, providing valuable insights into their psychological conditions. Emotions themselves are subjective responses arising from psychological and physiological stimuli, and they typically occur over short periods. Importantly, emotional expressions are generally universal across cultures, as established in early psychological studies [1]. Facial expressions result from the movement of facial muscles, which dynamically change in response to various emotional states. According to Paul Ekman's theory, there are seven basic emotional expressions that are universally recognized: anger, fear, happiness, sadness, surprise, disgust, and contempt [2].

In recent years, deep learning, particularly Convolutional Neural Networks (CNNs), has become a widely adopted approach for facial expression recognition [3,4,5]. CNNs are capable of automatically extracting hierarchical features from image data, starting from low-level patterns such as edges to high-level features like facial regions and emotional indicators [6]. This ability allows CNNs to perform robustly even when facial images vary in orientation or lighting conditions [7]. Despite their effectiveness, CNNs are not without limitations. These models often require large amounts of annotated training data to generalize well and are susceptible to overfitting. Moreover, training CNNs can be computationally intensive, often requiring high-performance GPUs. External factors such as cultural differences, age, and inconsistent lighting can also affect facial expression detection accuracy. Furthermore, CNNs are often criticized for their lack of interpretability, as their decision-making process is perceived as a "black box" [7].

Numerous studies have demonstrated that CNNs can achieve high accuracy in facial expression recognition tasks. For instance, a study using the ResNet model achieved an accuracy of 90.81% [8], while another version of ResNet reached 88% [9], and a model based on VGG architecture reported an accuracy of 87.71% [10]. These findings highlight the potential of CNN-based models for facial expression classification, although further investigation remains necessary, particularly in practical applications.

This study is motivated by the problem of how to accurately classify facial expressions using a robust and efficient deep learning model, despite challenges such as data variability, model complexity, and generalization ability. Therefore, the objective of this research is to implement and evaluate a CNN-based model for facial expression classification, using Ekman's universal emotional categories as the target output. The study aims to assess the model's accuracy and reliability in identifying expressions under varying facial conditions.

## 2. METHODOLOGY

### 2.1. Data Collection Method

This data collection method is collected by searching for various kinds of public datasets such as Kaggle. The type of data used is facial data in the form of images with 5 types of micro-expressions.

### 2.2. Design Method

We use the CNN VGG-16 model as the main model in this study. But before entering the model, there are several stages that must be gone through such as: data processing, Gridsearch, K-fold. These stages will help the researcher to modify some components in the model used to improve accuracy, and to evaluate the model using ablation testing to get to know or understand the structure of the model further. The following is the sequence and details of the design:

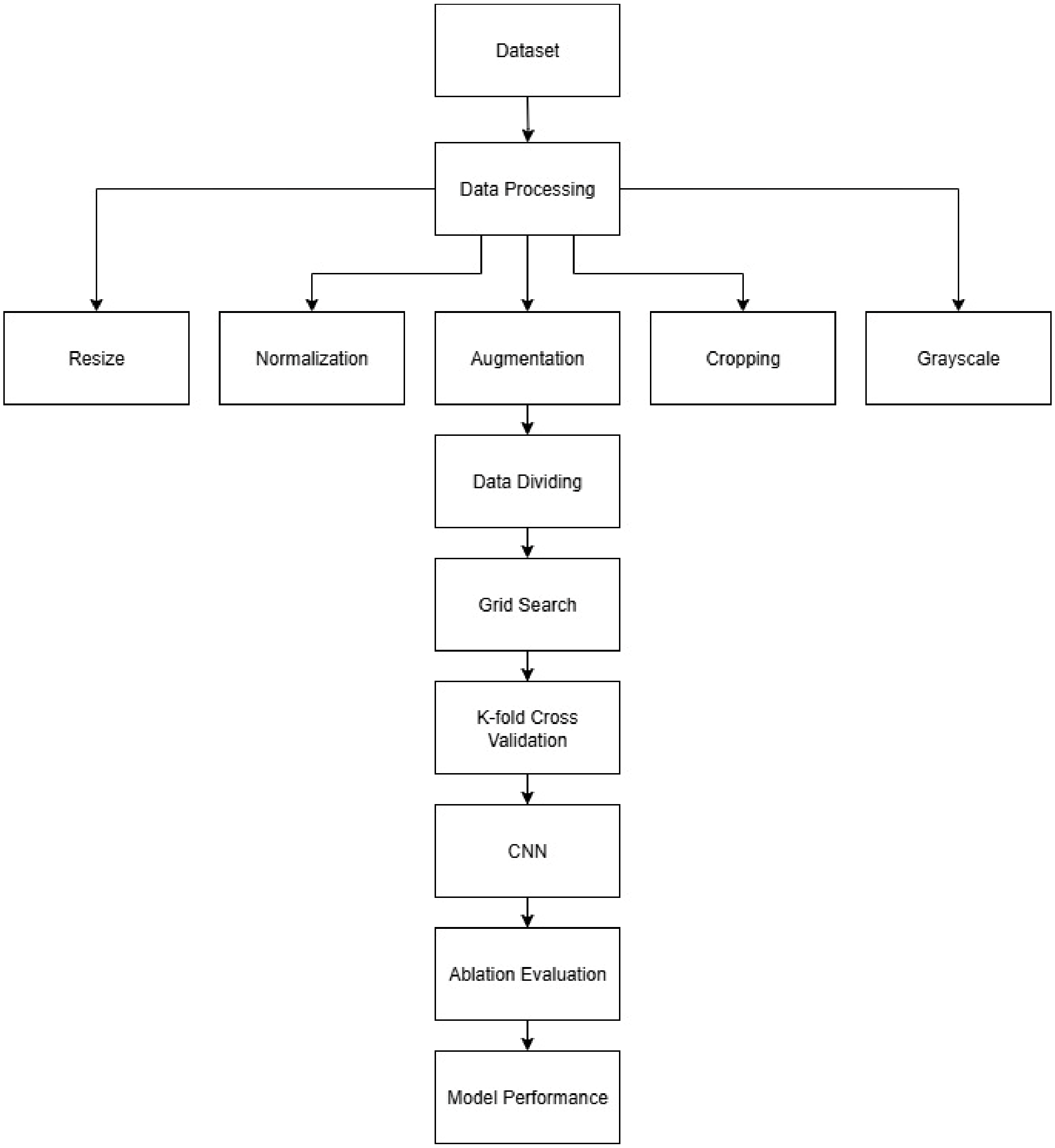

Figure 1. System process flow design

a. Dataset

The dataset in the design is a collection of data that the researcher has collected which contains images with 5 types of facial expression classes from the public dataset, namely *kaggle*.

b. Data Processing

In this stage, the data will be resized, formatted, data augmentation, cropped, and grayscaled (if necessary). This is to increase the level of accuracy when performing comparisons.

c. Data Division

This stage is to divide the dataset into 3 parts, namely: training data, validation data, and testing data. The distribution percentage is determined by the researcher.

d. Gridsearch

In this stage, the GridSearch algorithm is used to find the best parameter combination in a predetermined parameter space.

e. K-fold Cross Validation

In this step, the K-fold technique functions to evaluate the performance of the model by dividing the dataset into several subsets or folds.

f. CNN

Researchers chose VGG-16 because of its deep and structured architecture, which has proven to be very effective in extracting complex visual features from images. VGG-16 is able to capture various levels of information from facial images, from basic edges to more complex patterns such as the shape of the eyes, mouth, and eyebrows which are essential in determining facial expressions. The main advantage of VGG-16 is the use of a convolution layer with a small filter (3x3), which allows for more detailed and precise feature extraction without significantly increasing the number of parameters. For more complete steps are as follows:

1) Convolution layer

In this layer, a 3x3 size filter is used to perform a convolution operation on the input image. This function extracts basic features such as edges and textures. The ReLU (Rectified Linear Unit) activation function is applied to introduce non-linearity into the model, which helps the network learn from data more effectively.

2) Max Pooling

This layer reduces the dimension of the map feature by taking the maximum value in a 2x2 window. It serves to help significantly reduce the size of the data and make the model more efficient as well as help reduce overfitting.

3) Repetition

This convocation and max pooling layer repeats 5 times with an increasing number of filters that allows the network to capture more complex features and details from the image.

4) Full connect layer

This layer connects all neurons from the previous layer to the neurons in this layer. It helps to combine the extracted features to make classification decisions.

5) Softmax

The last layer with the number of neurons corresponding to the number of classes in the classification task (for example, 5 classes for 5 basic facial expressions). The function of this layer is used to generate probabilities for each class, enabling final classification of the input image.

g. Ablation repetition

In this stage, the model will be evaluated using ablation testing by modifying and comparing the model with and without some components.

h. Model Performance Results

In this case, the model can produce good performance to obtain higher accuracy compared to previous case studies.

## 3. RESULT AND DISCUSSION

The public datasets used are SMIC and SAMM, and based on the total data in the dataset after data processing, namely:

Table 1. Total amount of SMIC and SAMM data

| Class | Amount of data SMIC | Amount of data SAMM |
|---|---|---|
| Disgusted | 1837 | 4195 |
| Afraid | 2581 | 3420 |
| Happy | 2860 | 10.085 |
| Sad | 1400 | 2560 |
| Surprised | 3557 | 6300 |
| Total | 12.235 | 26.560 |

To train this VGG-16 model, data is needed for training, validation, and testing. Therefore, each dataset will be divided into 3 parts with the following percentages:

- SMIC :

Table 2. SMIC data distribution

| Class | Total | Training Data 70 % | Validation Data 15% | Testing Data 15% |
|---|---|---|---|---|
| Disgusted | 1837 | 1286 | 276 | 275 |
| Afraid | 2581 | 1807 | 387 | 387 |
| Happy | 2860 | 2002 | 429 | 429 |
| Sad | 1400 | 980 | 210 | 210 |
| Surprised | 3557 | 2490 | 533 | 534 |
| Total | 12.235 | 8565 | 1835 | 1835 |

- SAMM :

Table 3. SAMM data distribution

| Class | Total | Training Data 70 % | Validation Data 15% | Testing Data 15% |
|---|---|---|---|---|
| Disgusted | 4195 | 2937 | 629 | 629 |
| Afraid | 3420 | 2394 | 513 | 513 |
| Happy | 10.085 | 7059 | 1513 | 1513 |
| Sad | 2560 | 1792 | 384 | 384 |
| Surprised | 6300 | 4410 | 945 | 945 |
| Total | 26.560 | 18.592 | 3984 | 3984 |

To find the best combination of parameters with *Gridsearch*, several combinations were made with the following parameter restrictions:

Table 4. Searching for the best parameters with Gridsearch

| No | Filter | Learning rate | Dropout rate | Activation | Batch Size |
|---|---|---|---|---|---|
| 1 | 16 | 0.01 | 0.25 | relu | 8 |
| 2 | 32 | 0.001 | 0.5 | tanh | 16 |
| 3 | 64 | - | 0.75 | sigmoid | 32 |
| Best Parameter | 64 | 0.001 | 0.25 | relu | 32 |

Based on the results of the *training of* the VGG-16 model with the best combination for each dataset, the following results were obtained:

- SMIC:

Table 5. Accuracy results of the SMIC dataset

| K-Fold | Training | | Validasi | | pengujian | |
|---|---|---|---|---|---|---|
| | Accuration | loss | Accuration | loss | Accuration | loss |
| 1 | 0,9840 | 0,0507 | 0,9394 | 0,1923 | 0,9052 | 0,3460 |
| 2 | 0,9849 | 0,0463 | 0,9400 | 0,1847 | | |
| 3 | 0,9842 | 0,0478 | 0,9460 | 0,1649 | | |
| 4 | 0,9799 | 0,0570 | 0,9285 | 0,2179 | | |
| 5 | 0,9833 | 0,0546 | 0,9263 | 0,2572 | | |
| Average | 0,9832 | 0,0512 | 0,9360 | 0,2034 | | |

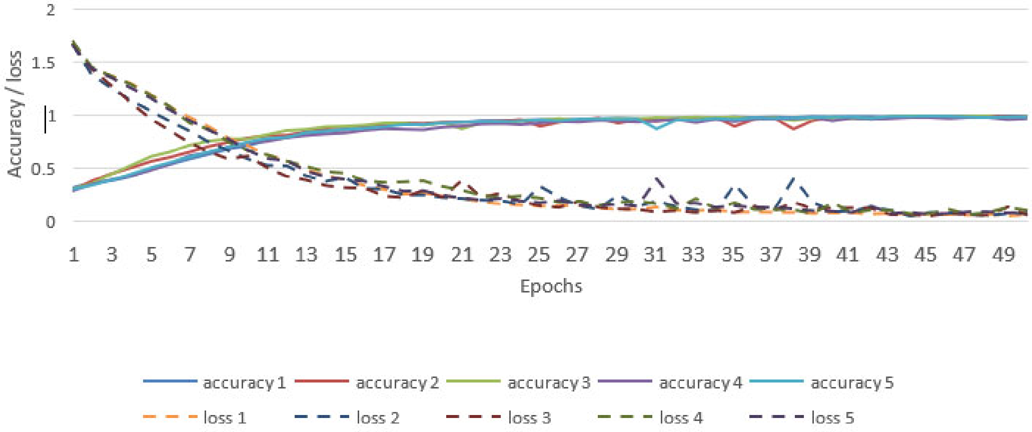

Figure 3. Accuracy and loss training of SMIC dataset

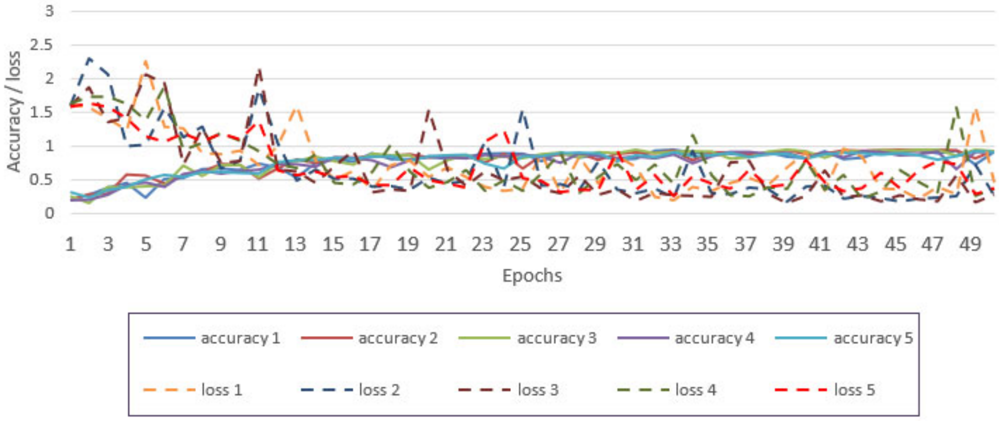

Figure 4. Validation of accuracy and loss of SMIC dataset

- SAMM :

Table 6. Accuracy results of the SAMM dataset

| K-Fold | *Training* | | Validasi | | Pengujian | |
|---|---|---|---|---|---|---|
| | Accuration | loss | Accuration | loss | Accuration | loss |
| 1 | 0,9960 | 0,0155 | 0,9829 | 0,0638 | 0,9804 | 0,0719 |
| 2 | 0,9949 | 0,0172 | 0,9839 | 0,0544 | | |
| 3 | 0,9958 | 0,0157 | 0,9844 | 0,0500 | | |
| 4 | 0,9952 | 0,0188 | 0,9859 | 0,0517 | | |
| 5 | 0,9945 | 0,0183 | 0,9786 | 0,0743 | | |
| Rata-rata | 0,9952 | 0,0171 | 0,9831 | 0,0588 | | |

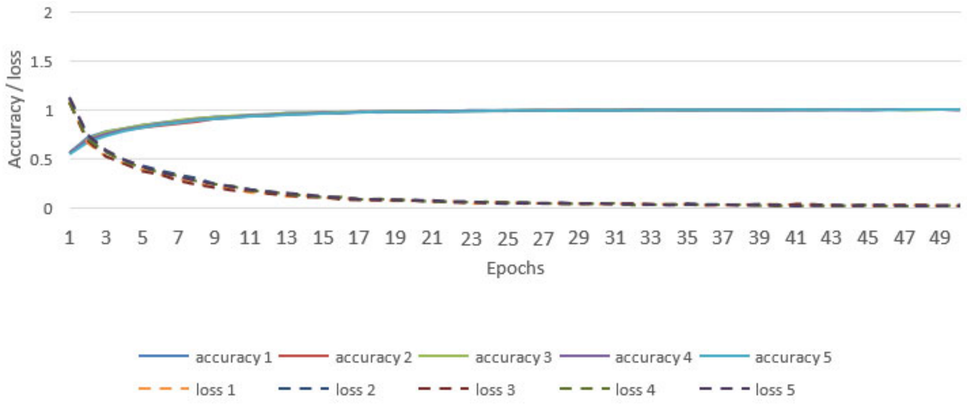

Figure 5. Accuracy and loss training of the SAMM dataset

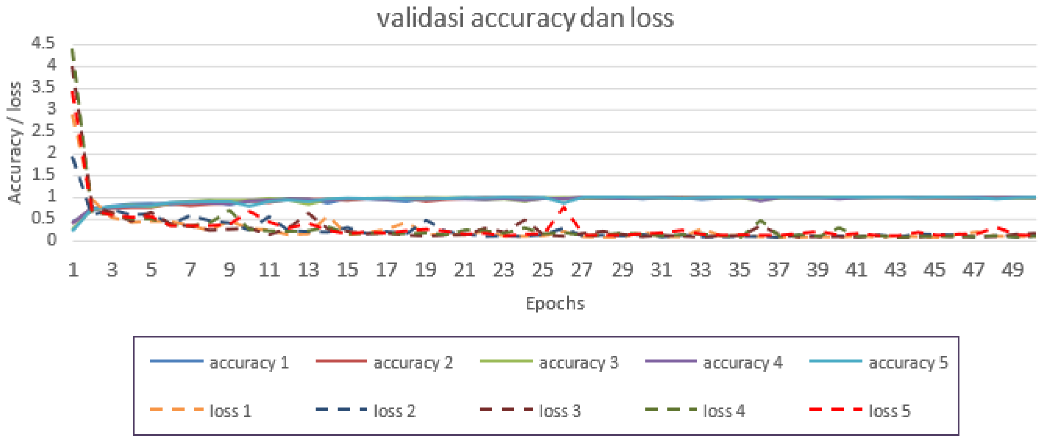

Figure 6. Validation of SAMM dataset accuracy and loss

Based on the model training process using the SMIC and SAMM datasets, the accuracy results from the SAMM dataset are higher than using the SMIC dataset. Because the accuracy results are higher using the SAMM dataset, the next step is to evaluate it using ablation testing to understand the model structure further using the SAMM dataset. The following is a table of ablation test results:

Table 7. Ablation test results on the model

| Model | Convolution Layer | Batch Normalization | Dropout | Validation Accuracy | Testing accuracy |
|---|---|---|---|---|---|
| Initial | 5 | Yes | Yes | 98,31% | 98,04% |
| Test 1 | 5 | Yes | Tidak ada | 98,43% | 98,22% |
| Test 2 | 5 | No | Yes | 84,67% | 84,69% |
| Test 3 | 5 | No | Tidak ada | 81,86% | 81,12% |
| Test 4 | 4 | Yes | Yes | 97,96% | 94,58% |
| Test 5 | 4 | No | Yes | 98,04% | 97,77% |
| Test 6 | 4 | Yes | No | 97,72% | 97,16% |
| Test 7 | 4 | No | No | 38,32% | 39,53% |
| Test 8 | 3 | Yes | Yes | 97,72% | 97,16% |
| Test 9 | 3 | No | Yes | 97,40% | 91,39% |
| Test 10 | 3 | Yes | No | 85,70% | 95,88% |
| Test 11 | 3 | No | No | 90,37% | 88,15% |

From the results obtained from the ablation experiment, it can be said that each component in the model structure affects the accuracy of the model. The following is an explanation of the results of ablation for each component of the model structure:

- Influence of the normalization batch:

In the initial model, the validation accuracy was 98.31% and the testing accuracy was 98.04%. When the normalization batch is eliminated, the accuracy tends to decrease as in test model 2. Even without dropout, models without normalization batches, as in test model 3, still experience a decrease in performance. However, there are also models that, without normalization batches, can increase the accuracy of the model, as in test 5. But the model in test 5 depends on the number of convulsions and dropouts.

- The effect of dropout

Removing dropout from the model can slightly improve model performance, as in the initial model and test model 1. However, in certain combinations such as in test model 6, removal results in poor performance. The use of dropout in models with few convulsion layers, such as in test model 8, can help maintain high accuracy.

- Effect of the number of convulsion layers

A decrease in the number of convolution layers from 5 to 3 as in test models 4 and 8 shows a decrease in model performance. When there are fewer layers and the normalization batch is eliminated, as well as the dropout, the model performance decreases significantly (test 7). But there is also a model structure with few layers but without normalization batch and dropout that can maintain the model performance even though the performance decreases slightly as in test 11.

## 4. CONCLUSION

The dataset does not fully affect the accuracy value. But the detail of the expression of an image can affect the accuracy of a model. The trial to find the best parameters using grid search can save time. K-fold cross validation helps the model to ensure that it can use all data for training and validation. And it can help the model work well on new, untrained data. With a modified model, it can produce high training accuracy and validation accuracy. The same goes for testing accuracy. After evaluating with ablation testing, we can conclude that each component in the model structure has its own important function. If any component is removed from the model structure, the performance of the model will decrease.

Before training the model, it is better to have a large total data of at least 5000 data per dataset. When training the model, it is better to use a high CPU or computer/laptop spec. Because training the model is very heavy. Avoid overfitting when training the model, so that the model can work on new data.

## REFERENCE

[1] E. Dalam, P. Lintas, B. Lilik, A. Budiono, and M. Masing, “INNOVATIVE: Volume 2 Nomor 1 Tahun 2022 Research & Learning in Primary Education.”

[2] P. Ekman, Universal facial expressions of emotion, vol. 8, California Mental Health Research Digest, Autumn 1970.

[3] Guntoro, A. L. S., Julianto, E., & Budiyanto, D. (2022). Pengenalan ekspresi wajah menggunakan Convolutional Neural Network. Jurnal Informatika Atma Jogja, 3(2), 155–160. https://doi.org/10.24002/jiaj.v3i2.6790.

[4] Elsheikh, R.A., Mohamed, M.A., Abou-Taleb, A.M. et al. Improved facial emotion recognition model based on a novel deep convolutional structure. Sci Rep 14, 29050 (2024). https://doi.org/10.1038/s41598-024-79167-8.

[5] Kopalidis, T.; Solachidis, V.; Vretos, N.; Daras, P. Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information 2024, 15, 135. https://doi.org/10.3390/info15030135.

[6] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, Joel J.P.C. Rodrigues, A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines, Alexandria Engineering Journal,

Volume 68, 2023, Pages 817-840, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2023.01.017.

[7] K. Grm, V. Struc, A. Artiges, M. Caron, and H. K. Ekenel, “Strengths and weaknesses of deep learning models for face recognition against image degradations,” *IET Biom*, vol. 7, no. 1, pp. 81–89, Jan. 2018, doi: 10.1049/iet-bmt.2017.0083.

[8] D. -H. Lee and J. -H. Yoo, "CNN Learning Strategy for Recognizing Facial Expressions," in IEEE Access, vol. 11, pp. 70865-70872, 2023, doi: 10.1109/ACCESS.2023.3294099.

[9] K. Zheng, D. Yang, J. Liu, and J. Cui, “Recognition of teacher’s facial expression intensity based on convolutional neural network and attention mechanism,” *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3046225.

[10] N. Zhou, R. Liang, and W. Shi, “A Lightweight Convolutional Neural Network for Real- Time Facial Expression Detection,” *IEEE Access*, vol. 9, pp. 5573–5584, 2021, doi: 10.1109/ACCESS.2020.3046715.