DOI: 10.6977/IJoSI.202502_9(1).0003
D.M. Ramadhan, H. Mubarok, etc./Int. J. Systematic Innovation, 9(1), 30-43 (2025)

Brightness augmentation implementation to evaluate performance classification of masked facial expressions based on the CNN model

Desty Mustika Ramadhan[1], Husni Mubarok[1], Rianto[1]*

[1]Department of Informatics, Siliwangi University, Tasikmalaya 46115, West Java, Indonesia

* Corresponding author E-mail: rianto@unsil.ac.id

(Received 07 June 2024; Final version received 20 September 2024; Accepted 21 October 2024)

Abstract

Deep learning methods with convolutional neural network (CNN) models have increasingly been applied to facial expression recognition. However, due to the recent pandemic, many individuals wear masks for work or health reasons, obstructing the complete visibility of their faces. This can impact social interactions, particularly in areas involving facial expression cues like the mouth. This study explores the application of CNNs in identifying facial expressions obscured by masks, focusing on the VGG16 and MobileNet architectures. Additionally, the research investigates the effects of data augmentation, including geometric and brightness augmentation, on the accuracy of facial expression classification. The findings indicate that the VGG16 architecture with cross-validation (VGG16-FLCV) outperforms MobileNet-FLCV in recognizing and classifying masked facial expressions. Data augmentation, particularly brightness augmentation, significantly enhances CNN model performance. For the VGG16-FLCV architecture, the brightness range (1.00, 1.25) yields the best accuracy, with a training accuracy of 81.73% and a validation accuracy of 70.71%. The most optimal brightness ranges for VGG16-FLCV are in the dark category (0.25, 0.50), (0.50, 0.75), and (0.75, 1.00), as well as the bright category (1.00, 1.25). Meanwhile, MobileNet-FLCV with brightness ranges (0.25, 0.50), (0.50, 0.75), (0.75, 1.00), (1.00, 1.25), and (1.25, 1.50) can be used as alternative brightness ranges without significant accuracy degradation. These findings provide valuable insights for improving the accuracy of masked facial expression recognition by applying appropriate data augmentation techniques.

Keywords: Brightness Augmentation, CNN, Cross-validation, Masked Facial Expressions

1. Introduction

Deep learning is a subset of machine learning. Therefore, it can be said that deep learning consists of a neural network with many layers and parameters. Most deep learning methods use a neural network (NN) architecture called deep NNs (Shinde & Shah, 2018).

The application of deep learning methods has gained prominence, particularly in facial expression recognition (FER). Most FER systems attempt to recognize expressions from a person's entire face. However, due to the pandemic in recent years, some people still wear masks to work or due to illness that requires them to wear a mask, which prevents their face from being fully visible (Castellano et al., 2021). The use of face masks has a negative effect. Psychologists report that it can confuse the reading of expressions, especially in the mouth region, which is very informative to help distinguish between expressions of sadness, disgust, fear, and surprise, thereby affecting social interactions (Yang et al., 2021).

There is research on classifying facial expressions with masks based on deep learning approaches. Grundmann et al. (2021) employed a strategy analytic approach using a multilevel regression model (logistic) to examine the effect of mask use on social judgment, investigate what facial cues are lacking when using masks to reduce expression recognition and explore how expression valence and associations related to the use of masks influence social judgment. Yang et al. (2021) proposed a facial expression type classification system for masked individuals based on a deep learning approach that applies convolutional neural network (CNN) models with MobileNetV2 and VGG19 architecture types. This research uses the M LFW-FER and M-KDDI-FER datasets, which have three types of facial expressions: positive, neutral, and negative. Hence, the masked facial expression classification process only considers facial expression classification with three expression categories. In addition, Castellano et al. (2021) used the CNN method with VGG16 and MobileNetV2 architecture types to recognize

emotions or expressions from the entire face and eye region of interest using the FER2013_cropped dataset with seven types of expressions, namely angry, disgust, fear, happy, neutral, sad, and surprise. The study investigated how well the FER system can recognize expressions, even when individuals wear face masks and expressions are often confused with others when the face is covered.

In addition, data augmentation techniques in deep learning models are often used in the image classification process that can handle data scarcity. Data augmentation can increase the accuracy value of the trained CNN model because it provides additional data, enhancing variations in the dataset used (Waheed et al., 2020). The type of data augmentation commonly used in research is geometry transformation (Kandel et al., 2022). Pei et al. (2019) applied data augmentation techniques in face recognition using CNN to address the issue of insufficient data samples. The types of data augmentation used in their research are geometric transformation, brightness augmentation, image translation, image rotation, image zoom, and filter operation. The results showed that applying the CNN method with data augmentation can achieve an accuracy of 86.3% higher than the principal component analysis or the local binary pattern histogram method. In addition, according to Kandel et al. (2022), applying data augmentation techniques for classification on histopathology, especially on the Invasive Ductal Carcinoma Dataset. They used two types of data augmentation, geometric and brightness augmentation, applying eight brightness scales using four CNN models, namely Resnet50, DenseNet121, InceptionV3, and Xception. The results showed that the application of geometric augmentation provides better accuracy than the application of brightness augmentation. In addition, the CNN model provides better results without the application of data augmentation techniques. Hence, it can be hypothesized that in the research conducted by Kandel et al. (2022), the application of brightness augmentation significantly reduced model performance when extreme values were used.

Based on the entire description of the work, gaps still pave the way for future research, primarily related to the type of expression used in Yang et al.'s research, which only has three categories: positive, negative, and neutral. The system was designed using CNN architecture, as shown by the research of Castellano et al., which uses VGG16 and MobileNetV2 architecture, while Yang et al. used VGG19 and MobileNetV2. The choice of architecture in these studies becomes a reference for applying transfer learning to the VGG16 and MobileNet architectures in classifying masked facial expressions. In addition, Pei et al. applied data augmentation, especially brightness augmentation, solely to face recognition datasets. Kandel et al. only applied eight brightness scales to histopathology datasets, particularly on the Invasive Ductal Carcinoma dataset.

Hence, this research is expected to provide a better understanding of masked facial expression recognition using deep learning techniques with the application of CNN architectures, namely VGG16 and MobileNet. In addition, data augmentation methods, such as geometric augmentation and brightness augmentation, will also be explored to evaluate their effect on the classification accuracy of masked facial expressions.

2. Related Work

Facial expression recognition systems mostly try to recognize expressions from a person's entire face. However, the pandemic has caused individuals to wear masks all the time, thereby causing their faces not to be fully visible. Currently, there is various research in classifying the types of facial expressions in masks. Table 1 shows the results of the related work and each proposed model's approach.

3. Methodology

This research methodology encompasses several key stages conducted systematically to achieve the research objectives. These stages include data collection, exploratory data analysis (EDA), data preparation, building the CNN model, model training, and model evaluation. The flow of these stages is illustrated in the flowchart shown in Fig. 1.

3.1. Data Collection

The research data from the Kaggle website and sample data for prediction tests were taken directly using a smartphone. The datasets used in this study include the MaskedDatasetFER, while 17 data samples were used for the prediction tests.

The explanation of the origin of the MaskedDatasetFER dataset is listed in the description on Kaggle (https://bit.ly/3UP0oRz), published in 2021. The MaskedDatasetFER dataset originated from the FER2013 dataset prepared by Pierre-Luc Carrier and Aaron Courville as part of an ongoing research project. A categorized dataset of masked individuals' facial expressions was created by artificially placing a face mask on the FER2013 dataset, forming a new dataset, MaskedDatasetFER.

MaskedDatasetFER consists of a dataset in the form of image files with image dimensions of 48 × 48 pixels with a red, green, and blue (RGB) image type. The total data from the dataset is 20,484 image data, consisting of 15,531 training data and 4,953 validation data. Both data have seven types of expressions or class labels: angry, disgusted, fearful, happy, sad,

Table 1. Related work

Author	Model	Description	Advantages	Limitations
Yang et al. (2021)	CNN model by applying VGG19 and MobileNetV2 architecture types.	Focus on facial expression recognition (FER) of three types of expressions on the people in masks.	The model effectively enhances facial expression recognition accuracy by focusing on uncovered areas and surpasses other mask-aware recognition methods.	This model only classified three emotional categories: positive, neutral, and negative, so its generalization is still limited.
Castellano et al. (2021)	Implement CNN model with VGG16 and MobileNetV2 architecture types and apply ADAM optimizer type.	Focus on introducing automatic expression of the expression face when wearing a mask.	This research has succeeded in developing an effective system for recognizing facial expressions only from the eye area.	This system has limitations in managing negative emotions, which often confuse expressions of sadness with anger or fear.
Pei et al. (2019)	The CNN model is used by applying VGG16 architecture types and the cross-validation method.	Focus on recognition face, recognition through learning deep learning using data augmentation based on experiment orthogonal.	The data augmentation method successfully increased the accuracy of facial recognition to 98.1% for class attendance.	The data collection process and the required orthogonal experiments can be challenging to implement.
Kandel et al. (2022)	Using CNN mode by applying Resnet50, DenseNet121, InceptionV3, and Xception.	Focus on brightness as an augmentative technique for image classification on an AGY dataset of invasive ductal carcinoma dataset.	Applying geometric augmentation techniques is more effective than brightness in improving CNN performance.	Brightness augmentation, especially at extreme values, can degrade model performance and not improve classification results.
Grundmann et al. (2021)	The multilevel regression (logistic) method was used with the nlminb optimizer, and the Glmer method was applied.	Focusing on face masks reduces expression recognition accuracy and perceived proximity.	This research reveals a significant impact of face mask use on the accuracy of emotion recognition and social judgment and provides essential insights for mask-related policy making.	Face masks have been shown to reduce the ability to classify emotional expressions accurately and may decrease feelings of closeness, particularly in older adults.
Cotter et al. (2020)	Implement CNN models with MobileNet, MobileEx, and ResNet architecture types.	Focus on introducing facial expression recognition on smartphones.	The proposed MobiExpressNet model has more than 5 times smaller size and FLOPs than the smallest MobileNet model, with 67.96% accuracy on the FER2013 dataset, making it very attractive for real-time smartphone applications.	The performance and accuracy of the ExpressNeT Mobile model have not been tested in real-world conditions on smartphone devices, which needs to be considered in further development.
Genc et al. (2020)	Using the Wizard of Oz	Focus on face mask design to reduce occlusion of facial expressions.	Electrochromic technology in smart masks enhances communication by displaying facial expressions.	The study was limited to a small sample and did not test the automated mechanism for real-world use.
Merghani et al. (2020)	Implement the FME algorithm and apply SMO, CASMEII, and SAMM methods.	Focus on creating a new adaptive mask for region-based facial micro-expression recognition by defining 14 new rois based on the most frequently used action units (au).	Region-based and adaptive mask methods for recognizing facial micro-expressions show promising accuracy, with competitive results compared to deep learning approaches on the SAMM dataset.	The accuracy of this method is still relatively low compared to other methods, and this research only evaluated two datasets without considering the potential combination with deep learning approaches in the future.

surprised, and neutral. The sample data from MaskedDatasetFER is provided in Fig. 2.

Meanwhile, the data samples used for the prediction tests have 17 image data samples with dimensions of 300 × 300 and are RGB image types. The sample data for the prediction test is shown in Fig. 3.

3.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is a process for exploring the dataset to be used. In this research, EDA includes two main aspects: data understanding and data visualization.

Table 1. *Cont'd*

Author	Model	Description	Advantages	Limitations
Yang et al. (2022)	Cross-attention-based and vision transformer models with CNN architecture types were used, including VGG19, MobileNet1, ResNet, and ViT, and the application of RAN, ACNN, and OADN methods.	Focus on Facial Expression Recognition based on Face Parsing and Vision Transformer.	The proposed method combines Transformer face parsing and vision models with a cross-attention mechanism to improve facial expression recognition accuracy with masks, surpassing the existing FER method.	This research has not tested the method in real-world contexts or broader scenarios beyond the dataset used.
Agrawal et al. (2020)	Using the CNN model and the FER2013 dataset.	Focus on studying the effect of kernel size and number of filters on facial expression recognition accuracy.	Presenting two new simple and effective CNN architectures, achieving 65% accuracy on the FER-2013 dataset.	This research is limited to the FER-2013 dataset, which may not be generalizable to other datasets or real-world applications.
Ding et al. (2020)	The CNN model was used with Resnet50, VGG16, and AffecNet architecture, and the OADN approach was applied.	Focus on occlusion-adaptive deep network for robust facial expression recognition.	This method improves the accuracy of expression recognition by handling occluded facial features and dividing the feature map into independent facial blocks for better robustness to occlusion.	This research requires high computational resources for model training and application. It may be less than optimal for very subtle or external expression variations not covered by the dataset.
Cheng et al. (2019)	Using model FSNet.	Focus on enhanced face segment or deep feature learning for face recognition.	FSENet enhances identity discrimination by exploiting local facial features through semantic segmentation and parsing maps while integrating global and local information.	The method may struggle with significant intra-personal variations or extreme facial condition changes not fully covered by the parsing maps.
Li et al. (2020)	Implementing the CNN model with VGG16 architecture and applying attention mechanisms and LBP approaches.	Focus on CNN-based attention mechanisms for facial expression recognition.	The proposed network integrates LBP features and an attention mechanism to enhance performance in facial expression recognition, demonstrating superior results on multiple datasets, including a newly collected one.	The method is currently limited to 2D images and does not address video data, 3D face datasets, or depth images, which could restrict its applicability.
Farkhod et al. (2022)	Using CNN model with Haar–Cascade classifier.	Focus on developing real-time landmark-based emotion recognition CNN for masked faces.	The proposed method achieves a high accuracy of 91.2% in image-based emotion recognition for masked faces by utilizing landmark-based features and a CNN model, showing effective performance compared to existing models.	Real-time emotion detection accuracy is lower due to biases and noise, such as image blurriness and poor lighting, and the method requires further development to improve performance under such conditions.

3.2.1. Data Understanding

The data understanding process in this research involves several crucial steps. Firstly, understanding the number of datasets in MaskedDatasetFER. Next, identify the types of classes present in the dataset. The following step involves evaluating the amount of data within each class in the dataset and understanding the data shape to ensure that the input dimensions are appropriate when building the CNN model. Since this research uses image data, verifying the RGB pixel values is necessary to ensure that the data used is indeed RGB image data.

3.2.2. Data Visualization

The data visualization process in this study involves two steps, including displaying the distribution of the MaskedDatasetFER dataset for both training and validation data and displaying images found in each class within the dataset.

3.3. Data Preparation

The data preparation stage is a crucial process involving the preprocessing of the dataset before the model training phase. In this research, the dataset was divided into two parts: training data and validation data. The data was split using various ratios, including 50:50, 60:40, 70:30, 80:20, and 90:10, to evaluate the model's performance based on different training and validation data proportions. Additionally, data augmentation techniques were applied, including brightness adjustment, rescaling, zooming, rotation, and other transformations to enhance the variety and quality of the training data. The brightness

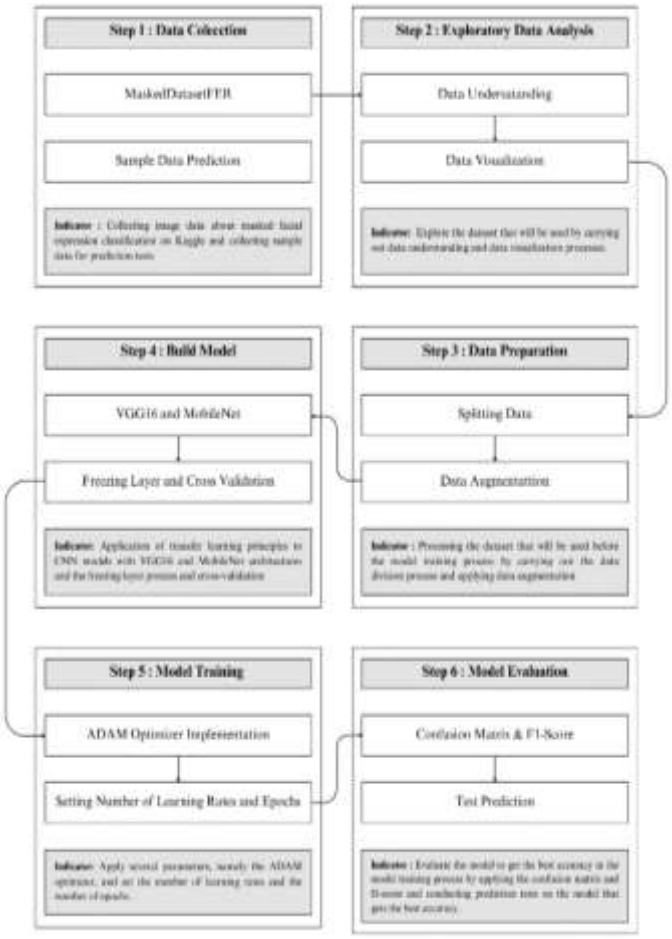

Fig. 1. Research flowchart.

Fig. 2. Sample MaskedDatasetFER.

Fig. 3. Sample prediction test.

augmentation specifically employed eight parameters, as used in the research by Kandel et al. (2022), with brightness ranges of 0.00–0.25, 0.25–0.50, 0.50–0.75, 0.75–1.00, 1.00–1.25, 1.25–1.50, 1.50–1.75, and 1.75–2.00. The results of these data augmentation techniques are illustrated in Fig. 4.

3.4. Building the CNN Model

Developing the CNN model in this research involves applying transfer learning to pre-trained architectures, namely VGG16 and MobileNet. The CNN architectures used in this study are VGG16-FL and MobileNet-FL, which apply the freezing layer technique to specific layers and modify the top layers. The detailed architectures of VGG16-FL and MobileNet-FL are illustrated in Fig. 5 and Fig. 6.

3.5. Model Training

At this stage, the CNN model that has been constructed is trained to achieve optimal performance. The training process involves the application of

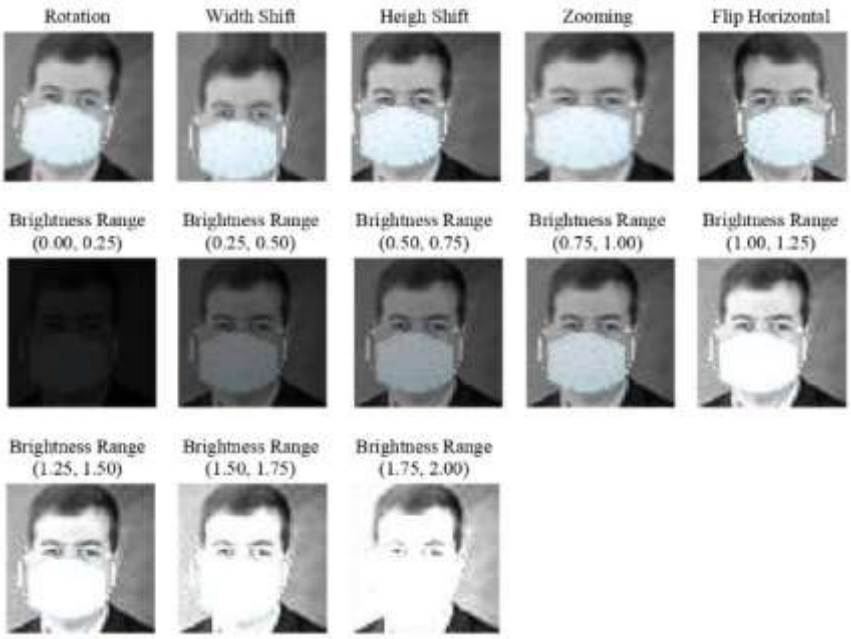

Fig. 4. Data augmentation.

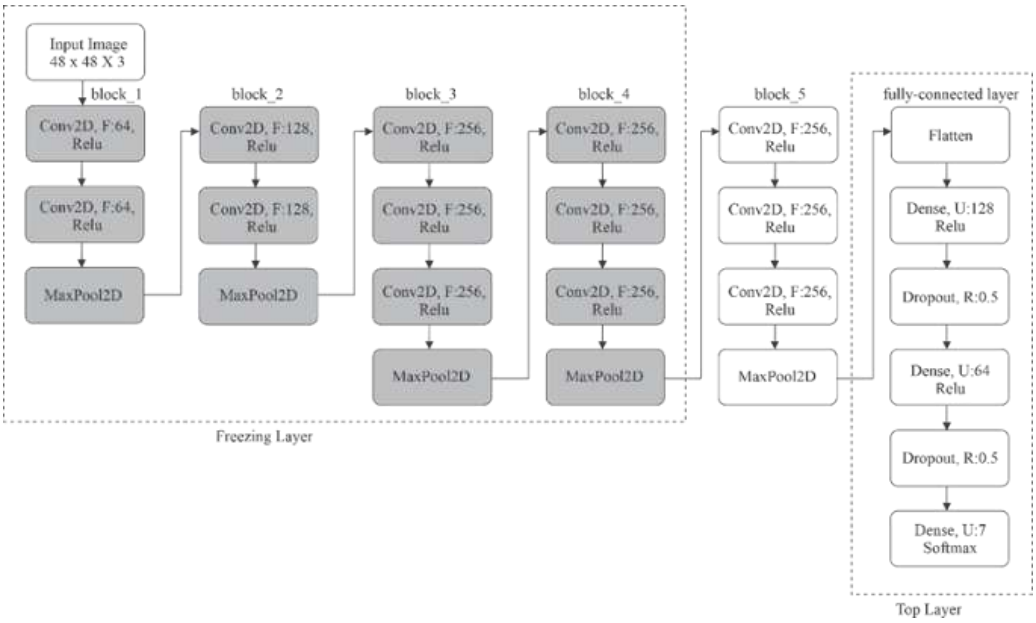

Fig. 5. VGG16-FL architecture.

various hyperparameters, including the ADAM optimizer. In this process, fine-tuning is performed by setting the learning rate at three levels: 0.001, 0.0001, and 0.00001. The number of epochs is also set to 100 to ensure the model can learn effectively from the available data.

The model training process applies the cross-validation method with k-fold values ranging from 2 to 10. This approach evaluates and improves the model's ability to generalize to the new data and address the imbalanced data issue. By comprehensively assessing the model's performance across various subsets of data, the model is expected to deliver consistent and accurate results, mainly when tested with uneven class distribution data.

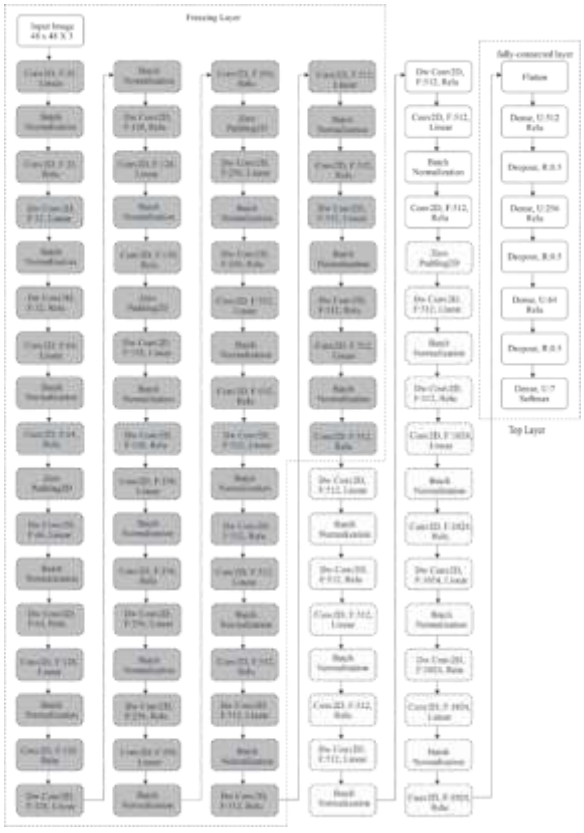

Fig. 6. MobileNet-FL architecture.

3.6. Model Evaluation

The system evaluation stage, which involves assessing the accuracy achieved in the research, calculates precision and recall values using the confusion matrix and F1-score methods. The calculations for accuracy, precision, recall, and F1-score based on the confusion matrix are presented in equations (1), (2), (3), and (4), according to Castellano et al. (2021).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

$$F1\ Score = \frac{(2\times Recall\times Precision\ )}{(Recall+Precision)} \quad (4)$$

T represents true, P represents positive, N represents negative, and F represents false.

Additionally, the evaluation process includes conducting predictive tests with actual data samples. This step is crucial to assess whether the developed model can accurately predict masked facial expressions using actual data, thereby validating the model's effectiveness in practical applications.

4. Results and Discussion

4.1. Analysis Model CNN

The VGG16-FL and MobileNet-FL architectures have the base pre-trained VGG16 and MobileNet models modified in the top layer, adding a flattened, dense, and dropout layer. Applying the dropout layer in both architectures is intended to reduce overfitting conditions, with dropouts of 0.2 and 0.5 applied.

Overfitting and underfitting conditions are two of the problems that cause inaccurate and suboptimal prediction results. Overfitting conditions can occur when a (NN that is overly dependent on the training set learns incorrect mappings that work well in the training set but perform poorly in the validation or testing set (Zhang et al., 2019). In addition, models trained with an unbalanced data set may become overfitted to training samples from underrepresented, resulting in poor generalization during test time (Li et al., 2021). There are several alternatives to handling models that experience overfitting conditions used in this study, including adding the application of data augmentation, which has been empirically proven to reduce overfitting with very high dimensional data by increasing the amount and variety of training data (Rice et al., 2020). Another way is the application of dropout by randomly discarding information targeting each hidden node of the NN during the training phase (Choe et al., 2019). A possible factor causing underfitting is that the NN architecture is too simple and has too few hidden layers or trainable parameters, making it not powerful enough to capture complex data characteristics (Zhang et al., 2019).

Based on previous experiments, the application of various data division ratios aims to obtain the best accuracy values. The results show that the two architectures used produce quite minimal accuracy

values and experience overfitting conditions. Data augmentation and dropout are applied to overcome the overfitting conditions, but in experiments without cross-validation, it is not effective enough to address overfitting. Therefore, the cross-validation method is applied to reduce overfitting in this research. In addition, overfitting can occur due to several factors, such as the dataset used having an imbalance of data between class labels for both training data and validation data. Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of a model and one of the methods used to prevent overfitting conditions (Berrar et al., 2018).

The application of the cross-validation method on both VGG16-FL and MobileNet-FL architectures in this study proved to be quite effective in handling overfitting. This is shown by the accuracy value obtained, which increased significantly from the experiment without cross-validation. The application of the number of k-folds in the cross-validation method also affects the accuracy value obtained, namely in the VGG16-FL cross-validation (VGG16-FLCV) architecture, resulting in the highest accuracy value at a value of 7-fold cross-validation. Meanwhile, in the MobileNet-FL cross-validation (MobileNet-FLCV) architecture, the highest accuracy value is at the 8-fold cross-validation value. Subsequently, brightness augmentation is applied to measure the effect on the accuracy value obtained. The graphs showing the accuracy and loss values on the VGG16 and MobileNet architectures and the application of cross-validation and brightness augmentation methods are illustrated in Fig. 7 to Fig. 14.

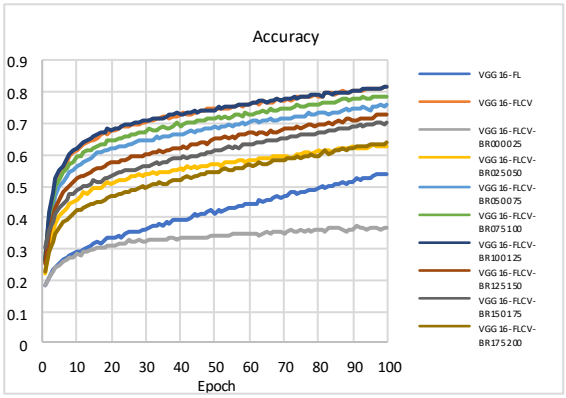

Fig. 7. Training accuracy with VGG16-FL.

Fig. 7 shows that the VGG16 architecture applying the cross-validation method has a higher training accuracy (VGG16-FLCV) than the VGG16 architecture without the cross-validation method. In addition, the graph shows that the application of brightness augmentation can increase the training accuracy value obtained precisely in the brightness range (1.00, 1.25), which obtained a training accuracy value of 81.73%. Meanwhile, the application of the brightness range (0.00, 0.25) has the lowest accuracy value compared to other experiments, with a training accuracy value of 36.78%. Hence, it can be concluded that applying the brightness range (0.00, 0.25) can significantly reduce the training accuracy value. The comparison graph of the loss value between models that do not apply the cross-validation method with models that apply the cross-validation method and the comparison of each application of brightness augmentation on the VGG16-FLCV architecture is illustrated in Fig. 8.

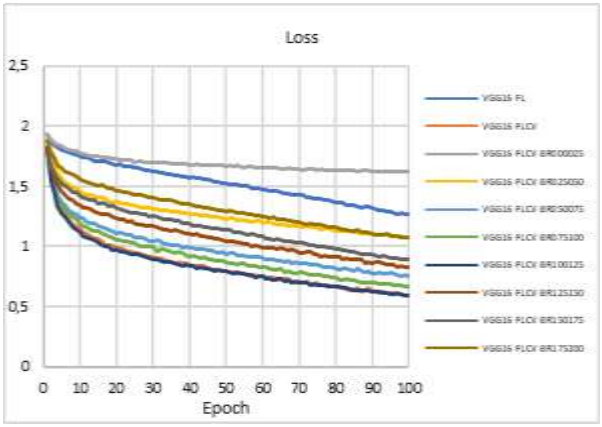

Fig. 8. Training loss with VGG16-FL.

Meanwhile, the training loss value graph in Fig. 8 shows that the smallest training loss value is found in the VGG16 architecture that applies the cross-validation method and brightness augmentation with the brightness range (1.00, 1.25). The highest training loss value listed in the graph is on the architecture applying brightness range (0.00, 0.25). In addition, the validation accuracy and loss graphs with the VGG16 architecture can also be seen in Fig. 9 and Fig. 10.

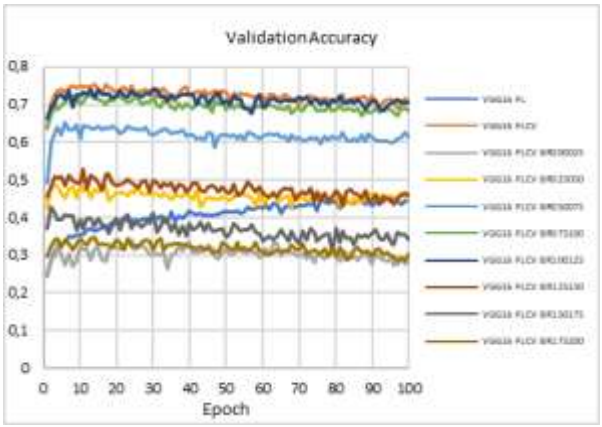

Fig. 9. Validation accuracy with VGG16-FL.

Fig. 9 shows a comparison graph of validation accuracy illustrating differences in the previous training accuracy graph. This validation accuracy shows relatively stable results, with only minor

fluctuations in the form of increases or decreases in validation accuracy values. However, the highest validation accuracy value is observed in experiments applying the cross-validation method (VGG16-FLCV) and in experiments applying brightness augmentation with brightness range (1.00, 1.25). The validation accuracy value in the VGG16-FLCV experiment is 70.61% and 70.71% in the brightness range (1.00, 1.25) experiment. In addition, the validation accuracy value that has the lowest value is in the experiment applying brightness augmentation (0.00, 0.25) with the resulting validation accuracy value of 28.26%.

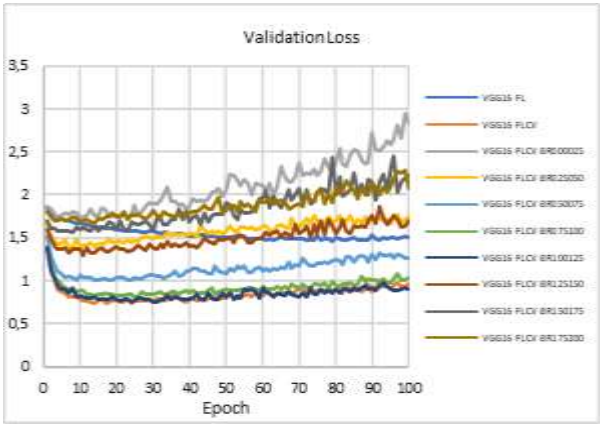

Fig. 10. Validation loss with VGG16-FL.

The graph in Fig. 10 shows that the experiment that applies the cross-validation method and the experiment that applies the brightness range (1.00, 1.25) have relatively low validation loss values compared to the validation loss value in other experiments. The experiment applying the brightness range (0.00, 0.25) has the highest validation loss value compared to other experiments.

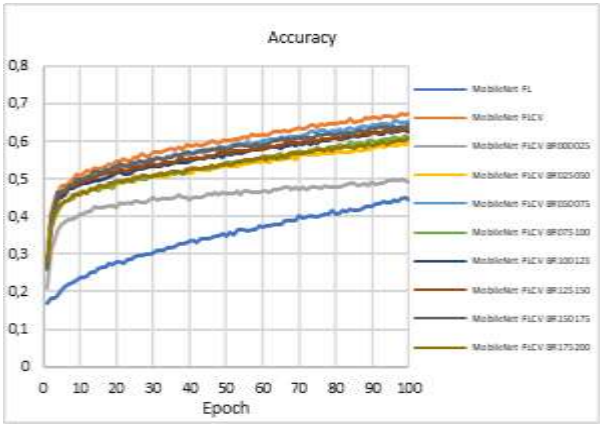

Fig. 11. Training accuracy with MobileNet-FL.

Based on the graph in Fig. 11, the MobileNet architecture applying the cross-validation method (MobileNet-FLCV) has a higher training accuracy compared to the MobileNet architecture without the cross-validation method. In addition, the graph shows that the application of brightness augmentation with the MobileNet architecture applying the brightness range (0.00, 0.25) can significantly reduce the accuracy value. In addition, the training accuracy value with the highest value is the experiment applying the cross-validation method without employing brightness augmentation, obtaining a training accuracy value of 67.47%. The lowest training accuracy value was found in the MobileNet architecture experiment without cross-validation, resulting in a training accuracy value of 44.46%.

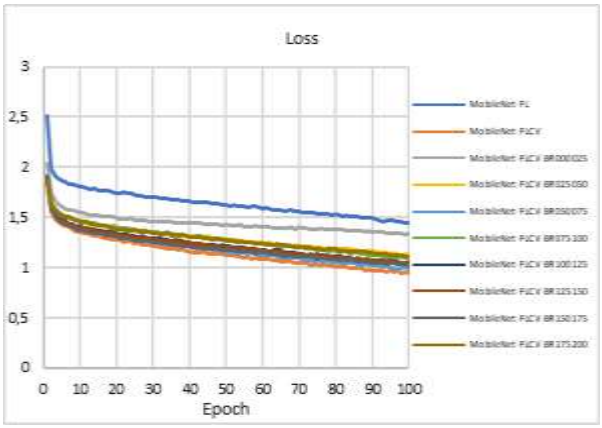

Fig. 12. Training loss with MobileNet-FL.

Meanwhile, the training loss value graph in Fig. 12 shows that the lowest training loss value is the MobileNet architecture applying the cross-validation method without brightness augmentation, and the highest training loss value is the experiment without applying the cross-validation method.

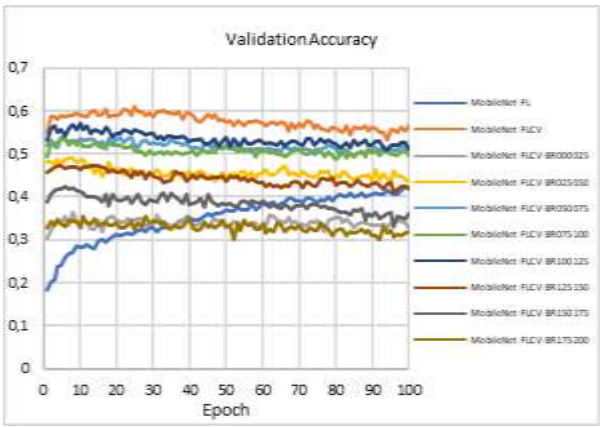

Fig. 13. Validation accuracy with MobileNet-FL.

Fig. 13 shows a comparison graph of validation accuracy showing differences in the previous training accuracy graph. This validation accuracy graph shows relatively constant results, indicating similarities in the experiments using the VGG16 architecture. Hence, it can be said that the changes that occur in the graph, both in the form of an increase or decrease in the validation accuracy value that occurs, are not too

significant. However, the highest validation accuracy value is the experiment that applies the cross-validation method or MobileNet-FLCV with a validation accuracy value of 56.37%. Meanwhile, the lowest validation accuracy value is found in the experiment applying the brightness range (1.75, 2.00) with a validation accuracy value of 31.80%.

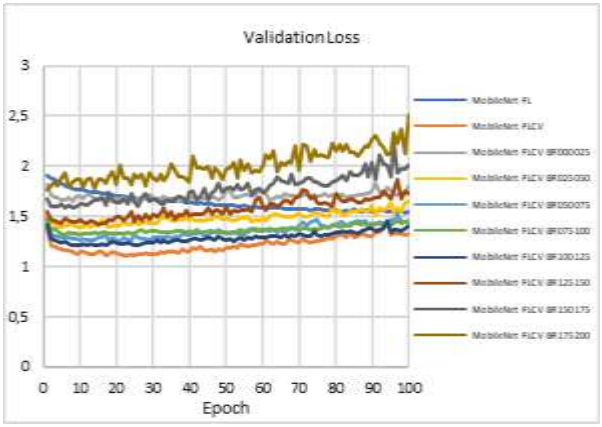

Fig. 14. Validation Loss with MobileNet-FL.

The graph in Fig. 14 shows that the experiment that applies the cross-validation method has a relatively small validation loss value compared to the validation loss value in other experiments, and the highest validation loss value is in the experiment with the brightness range (1.75, 2.00).

Based on the graphical results in Fig. 7 to Fig. 14, it can be said that the application of brightness augmentation can reduce and increase the accuracy values obtained. The application of brightness augmentation in the VGG16-FLCV architecture that can be applied in the model training process on the classification of masked facial expressions is the brightness range (0.25, 0.50), (0.50, 0.75), (0.75, 1.00), and (1.00, 0.25) ranges do not experience a very significant decrease in accuracy value. The application of brightness augmentation with a brightness range (1.00, 0.25) can increase the accuracy value from the accuracy value results that do not apply brightness augmentation in the VGG16-FLCV architecture. The accuracy values obtained from the application of brightness range (1.00, 0.25) with VGG16-FLCV architecture are 81.73% (training accuracy) and 70.71% (validation accuracy). Meanwhile, the model evaluation process is also carried out on the VGG16-FLCV architecture that applies brightness augmentation. Application of brightness range (0.00, 0.25) produces lower values compared to other experiments, namely with a precision value of 36.35%, recall of 34.03%, and F1-score of 34.81%. In addition, the application of brightness augmentation produces the highest value of prediction accuracy, precision, recall, and F1-score, namely in the use of brightness range (0.75, 1.00) in 7-fold cross-validation with a value of precision 74.51%, recall 72.38%, and F1-score 73.22%, and the use of brightness range (1.00, 1.25) in 7-fold cross-validation with precision 76.23%, recall 74.16%, and F1-score 74.97%.

Meanwhile, the application of brightness augmentation in the MobileNet-FLCV architecture mostly decreases the accuracy value of the experiment that does not apply brightness augmentation. The application of brightness augmentation that has a fairly high accuracy value among other brightness augmentation applications is the brightness range (1.00, 1.25), which obtained a training accuracy value of 62.66% and a validation accuracy value of 51.21%. So, it can be said that the application of brightness augmentation in the MobileNet-FLCV architecture that can be applied in the model training process on the classification of masked facial expressions is in the brightness range (0.25, 0.50), brightness range (0.50, 0.75), brightness range (0.75, 1.00), brightness range (1.00, 0.25) and brightness range (1.25, 1.50). Meanwhile, the model evaluation results with the application of brightness augmentation and cross-validation show that the Mobilenet architecture can have smaller precision, recall, and F1-score values compared to the application of the VGG16 architecture. The precision, recall, and f1-score values are quite high compared to the other experiments in the application of MobileNet-FLCV architecture found in the experiment without applying brightness augmentation with 8-fold cross-validation, which is 63.58% precision, 60.93% recall, and 61.26% f1-score.

This study compares several models based on five performance metrics: precision, recall, F1-score, accuracy, and validation accuracy. The models include various VGG16 and MobileNet variants, each evaluated under different configurations, as shown in Table 2.

The results indicate that the application of cross-validation significantly improves the performance of both VGG16 and MobileNet models. However, the effects of brightness augmentation show variability, suggesting that careful adjustment of augmentation parameters is essential to optimize model performance.

Based on research conducted by Kandel et al. (2022), the application of geometric augmentation provides a better accuracy value compared to the application of brightness augmentation. In addition, the CNN model, without applying data augmentation techniques, gives better results than the application of brightness augmentation. However, in this study, the application of brightness augmentation can increase the resulting accuracy value, namely in the application of VGG16-FLCV architecture with the brightness range (1.00, 1.25) in the process of masked facial expression classification. This research proposes several brightness range parameters that can be used for the masked facial expression classification process for both darkness and brightness categories in each architecture used, namely VGG16-FLCV and MobileNet-FLCV. In addition, there are similarities

Table 2. Comparison of value precision, recall, f1-score, accuracy, and validation accuracy.

Techniques used	Precision	Recall	F1-Score	Accuracy	Validation accuracy
VGG16-FL	0.1766	0.1689	0.1713	0.5413	0.4454
VGG16-FLCV	0.7716	0.7535	0.7617	0.8165	0.7061
VGG16-FLCV-BR000025	0.3635	0.3403	0.3481	0.3678	0.2826
VGG16-FLCV-BR025050	0.5272	0.4835	0.4950	0.6303	0.4491
VGG16-FLCV-BR050075	0.6777	0.6527	0.6604	0.7592	0.612
VGG16-FLCV-BR075100	0.7451	0.7238	0.7322	0.7884	0.682
VGG16-FLCV-BR100125	0.7623	0.7416	0.7497	0.8173	0.7071
VGG16-FLCV-BR125150	0.5496	0.5280	0.5344	0.7270	0.4573
VGG16-FLCV-BR150175	0.5395	0.4289	0.4571	0.7040	0.3401
VGG16-FLCV-BR175200	0.4211	0.3468	0.3608	0.6372	0.3028
MobileNet-FL	0.2169	0.1723	0.1843	0,4446	0.4191
MobileNet -FLCV	0.6586	0.6351	0.6440	0.6875	0.5512
MobileNet-FLCV-BR000025	0.4062	0.3941	0.3959	0.5214	0.3602
MobileNet-FLCV-BR025050	0.5107	0.4968	0.4966	0.6053	0.4437
MobileNet-FLCV-BR050075	0.6117	0.5843	0.5936	0.6699	0.5094
MobileNet-FLCV-BR075100	0.5992	0.5500	0.5655	0.6102	0.4941
MobileNet-FLCV-BR100125	0.5961	0.5746	0.5798	0.6443	0.5230
MobileNet-FLCV-BR125150	0.5824	0.4636	0.4887	0.6298	0.4199
MobileNet-FLCV-BR150175	0.4908	0.4128	0.4253	0.6432	0.3734
MobileNet-FLCV-BR175200	0.5021	0.3675	0.3926	0.6089	0.3102

between masked facial expression classification research contained in Yang et al.'s (2021) research and Castellano et al.'s (2021) research in terms of using CNN architecture, which is a reference in this research. However, there are differences in this study, namely, in this study, there is an additional method to optimize the accuracy value obtained by adding the cross-validation method to the VGG16 and MobileNet architectures.

4.2. Analysis Prediction Test

The prediction test process is carried out using prediction sample data to test whether the model can predict masked facial expressions with real data. The prediction test process is carried out using a model with the application of the VGG16-FLCV architecture in 7-fold cross-validation and MobileNet-FLCV 8-fold cross-validation and with the application of brightness augmentation. The results of the prediction test process can be seen in Fig. 15 to Fig. 18.

Fig. 15 shows the results of the types of expressions produced in the 17 samples used in the prediction test process on the VGG16-FLCV architecture based on the techniques used. In addition, based on this figure, it shows that the model that can predict the most types of expressions is the model that applies the brightness range (0.50, 0.75), which in 17 samples predicts angry, neutral sad, happy, fear, and surprise expressions. Meanwhile, the model that can predict the least types of expressions is the model that applies the brightness range (1.50, 1.75).

The prediction test was carried out on 17 data samples; each sample has a difference in predicting the type of facial expression in the application of the VGG16-FLCV architecture. The graph that shows the results of the expression type prediction test for each sample can be seen in Fig. 16.

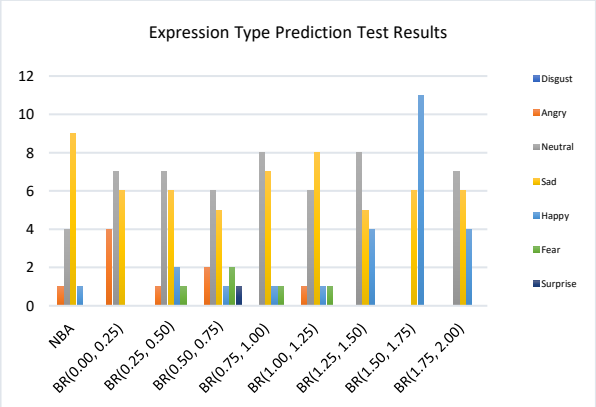

Fig. 15. Expression type prediction test results VGG16-FLCV.

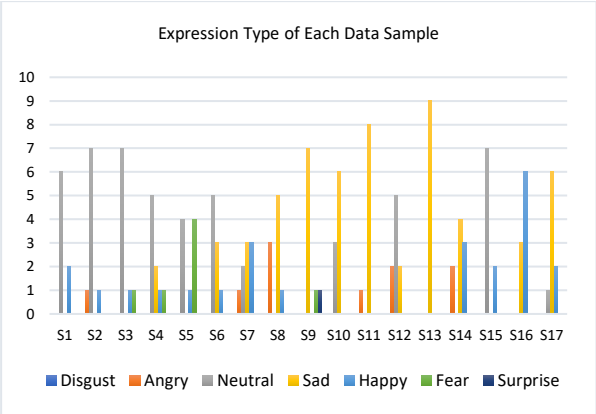

Fig. 16. Expression type of each data sample VGG16-FLCV.

Fig. 16 shows the number of facial expression prediction results for each data sample used in the prediction test with the VGG16-FLCV architecture. Data samples that have the highest number of expression predictions are the fourth (S4) and seventh (S7) data samples. The fourth data sample (S4) is predicted to have neutral, sad, happy, and fear expression types. Meanwhile, the data sample that has the least number of predictions, with one type of expression, is the 13th data sample (S13), which is predicted to have a sad expression type.

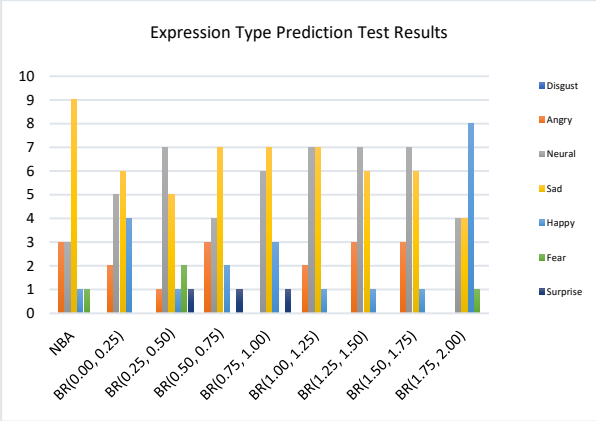

Fig. 17. Expression Type Prediction Test Results MobileNetFLCV.

Fig. 17 shows the results of expression types generated in 17 samples used in the prediction test process with the MobileNet-FLCV architecture based on the techniques used. In addition, based on the figure, it shows that the model that can predict the most types of expressions is the model that applies the brightness range (0.25, 0.50), which in 17 samples predicted angry, neutral sad, happy, scared, and surprised expressions. The graph that shows the prediction test results for each sample can be seen in Fig. 18.

Fig. 18 shows the number of facial expression prediction results for each data sample used in the prediction test with MobileNet-FLCV architecture. Data samples that have four types of expression predictions are in the sixth (S6), seventh (S7), 15th (S15), and 17th (S17) data samples. In addition, data samples that predicted only one type of expression were in the third data sample (S3), the eighth data sample (S8), the 12th data sample (S12), and the 13th data sample (S13).

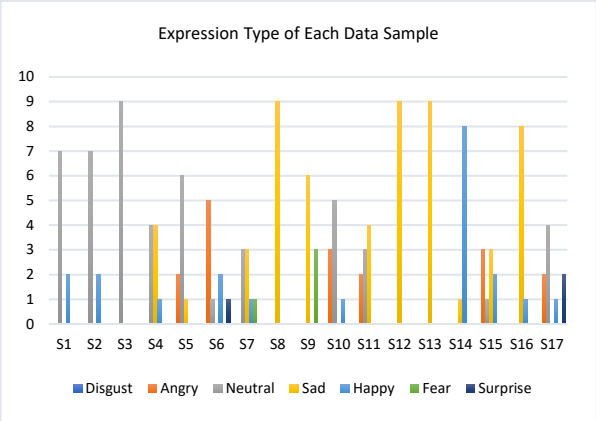

Fig. 18. Expression type of each data sample MobileNet-FLCV.

5. Conclusions

This research has successfully applied deep learning techniques using CNN architectures, specifically VGG16 and MobileNet, for masked facial expression classification. This research makes an important contribution to overcoming the challenges of facial expression recognition when wearing masks makes it difficult in social interactions. The results show that the use of VGG16 architecture with cross-validation method (VGG16-FLCV) provides better performance than MobileNet-FLCV architecture in recognizing and classifying masked facial expressions. This shows that in masked facial expression classification applications, VGG16 is superior to MobileNet. The application of data augmentation methods, such as geometric augmentation and brightness augmentation, has helped to improve the performance of CNN models. However, it is important to choose an appropriate brightness range value to obtain optimal results. The experimental results show that on the VGG16-FLCV architecture, the brightness range (1.00, 1.25) provides the best accuracy with a training accuracy of 81.73% and 70.71% validation accuracy.

The application of brightness augmentation on the MobileNet-FLCV architecture does not provide comparable performance to VGG16-FLCV. From the results of this study, it is found that the optimal application of brightness range on VGG16-FLCV architecture is in the darkness category with the brightness ranges (0.25, 0.50), (0.50, 0.75) and (0.75,

1.00) and in the brightness category with brightness range (1.00, 1.25). In addition, this study found that the MobileNet-FLCV architecture with brightness ranges (0.25, 0.50), (0.50, 0.75), (0.75, 1.00), (1.00, 0.25), and (1.25, 1.50) can be used as an alternative brightness range without experiencing a significant decrease in accuracy. Therefore, the results of this study provide a reference for the selection of the right brightness range in the application of data augmentation in CNN models, especially in the context of masked facial expression classification. This information is potentially useful for the development of more effective masked facial expression recognition technology and to support social interactions in pandemics or environments with extensive use of masks.

This research possesses several weaknesses and shortcomings. Therefore, the following suggestions can be used as a reference for further research and development, including applying other types of pre-trained model architectures in implementing transfer learning principles for the classification or prediction of masked facial expression types. Age parameters can be added to the process of predicting the type of masked facial expression to predict the type of masked facial expression and age of masked individuals. The development of the CNN model in predicting masked facial expressions is needed so that the process of detecting masked facial expressions can be done in real-time or be used for developing applications to detect masked facial expressions.

References

Agrawal, A. & Mittal, N. (2019). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 2. https://doi.org/10.1007/s00371-019-01630-9

Berrar, D. (2018). Cross-validation. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1–3(April), 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Castellano, G., De Carolis, B. & Macchiarulo, N. (2021). Automatic emotion recognition from facial expressions when wearing a mask. ACM International Conference Proceeding Series. https://doi.org/10.1145/3464385.3464730

Cheng, X., Lu, J., Member, S. & Yuan, B. (n.d.). Face Segmentor-Enhanced Deep Feature Learning for Face Recognition. 1–14.

Choe, J. & Shim, H. (n.d.). Attention-based Dropout Layer for Weakly Supervised Object Localization. 2219–2228.

Cotter, S.F. (n.d.). MobiExpressNet: A Deep Learning Network for Face Expression Recognition on Smart Phones. 2020 IEEE International Conference on Consumer Electronics (ICCE), 1–4.

Ding, H., Zhou, P. & Chellappa, R. (2020). Occlusion-adaptive deep network for robust facial expression recognition. IJCB 2020 - IEEE/IAPR International Joint Conference on Biometrics. https://doi.org/10.1109/IJCB48548.2020.9304923

Farkhod, A. & Abdusalomov, A.B. (2022). Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces.

Genç, Ç., Colley, A., Löchtefeld, M. & Häkkilä, J. (2020). Face mask design to mitigate facial expression occlusion. Proceedings - International Symposium on Wearable Computers, ISWC, 40–44. https://doi.org/10.1145/3410531.3414303.

Grundmann, F., Epstude, K. & Id, S.S. (2021). Face masks reduce emotion-recognition accuracy and perceived closeness. 1–18. https://doi.org/10.1371/journal.pone.0249792.

Kandel, I., Castelli, M., & Manzoni, L. (2022). Brightness as an Augmentation Technique for Image Classification. Emerging Science Journal, 6(4), 881–892. https://doi.org/10.28991/ESJ-2022-06-04-015.

Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention Mechanism-based CNN for Facial Expression. Neurocomputing. https://doi.org/10.1016/j.neucom.2020.06.014

Li, Z., Kamnitsas, K., & Glocker, B. (2021). Analyzing Overfitting Under Class Imbalance in Neural Networks for Image Segmentation. 40(3), 1065–1077.

Merghani, W., & Yap, M. H. (n.d.). Adaptive Mask for Region-based Facial Micro-Expression Recognition. 2–7.

Pei, Z., Xu, H., Zhang, Y., Guo, M., & Yang, Y. (2019). Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments. 1–16. https://doi.org/10.3390/electronics8101088.

Rice, L., Wong, E., & Kolter, J. Z. (2020). Overfitting in adversarially robust deep learning. 37th International Conference on Machine Learning,

ICML 2020, PartF16814, 8049–8074.

Shinde, P.P. & Shah S. (2018). A Review of Machine Learning and Deep Learning Applications. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F. & Pinheiro, P.R. (2020). CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. IEEE Access, 8, 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762.

Yang, B. (2021). Face Mask Aware Robust Facial Expression Recognition. 240–244.

Yang, B., Wu, J., Ikeda, K., Hattori, G., Sugano, M., Iwasawa, Y., & Matsuo, Y. (2022). Face-mask-aware Facial Expression Recognition based on Face Parsing and Vision Transformer. Pattern Recognition Letters, 164, 173–182. https://doi.org/10.1016/j.patrec.2022.11.004.

Zhang, H., Zhang, L., & Jiang, Y. (n.d.). Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems. 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 1–6.

AUTHOR BIOGRAPHIES

Desty Mustika Ramadhan completed her undergraduate education in informatics at Siliwangi University. During her study period, she actively participated in the independent study program (MSIB) at the independent campus at Bangkit Academy, which took her on the machine learning path.

Husni Mubarok completed his master's education at the Bandung Institute of Technology in 2011. He is currently a researcher and lecturer in informatics engineering at Silwangi University, Tasikmalaya. Research areas include Artificial Intelligence, Machine Learning, Recognition Systems, Intelligent Systems, and Pattern Recognition.

Rianto completed his master's education at the Bandung Institute of Technology in 2015. He is currently a researcher and lecturer in informatics engineering at Silwangi University, Tasikmalaya. Research areas include Software Engineering, Artificial Intelligence, Data Mining, Data Processing, Text Mining, Sentiment Analysis, Information Science.