Effect of Data Augmentation Levels on MobileNetV2 Performance for Facial Expression Recognition Using FER-2013 Dataset

Ramlan Marbun, Muhammad Syahputra Novelan, Muhammad Irfan Sarif

Abstract

This research is conducted to assess how data augmentation techniques influence the performance of deep learning models in the task of facial expression classification. A critical issue commonly encountered in facial image analysis is the scarcity of available datasets, which frequently results in overfitting and limits the model's ability to generalize effectively to unseen data. In response to this limitation, the study implements two distinct levels of augmentation, referred to as light augmentation and complex augmentation, and evaluates their performance against a baseline condition where no augmentation is applied. The model employed in this investigation is MobileNetV2, trained using the FER-2013 dataset that contains seven distinct emotion categories. The findings from the experimental evaluation indicate that the application of light augmentation yields the highest validation accuracy at 22.7%, outperforming both the no-augmentation scenario (10.8%) and the complex augmentation approach (8.3%). Although the use of complex augmentation results in lower loss values, it does not translate into improved classification accuracy. This outcome suggests that overly intensive augmentation strategies may hinder the model's capacity to effectively learn and extract meaningful features. Overall, these results highlight the necessity of carefully determining appropriate augmentation methods to enhance the performance of deep learning models.

Keywords: *Deep Learning, Data Augmentation, MobileNetV2, Facial Expression Recognition*

Ramlan Marbun[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: ramlanlumbangaol90@gmail.com[1]

Muhammad Syahputra Novelan[2], Muhammad Irfan Sarif[3]
[2,3]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: putranovelan@dosen.pancabudi.ac.id[2], irfanberbagi@gmail.com[3]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Learning concentration represents a crucial determinant in shaping students' success throughout the educational process. One methodological approach that can be employed to assess concentration levels involves analyzing facial expressions as observable behavioral indicators. With the rapid advancement of artificial intelligence, particularly within the domain of deep learning, such facial expression analysis can now be conducted automatically through the utilization of Convolutional Neural Network (CNN) architectures. These models enable the extraction and interpretation of visual features in a systematic and scalable manner, thereby supporting more objective assessments of student engagement. As a result, integrating CNN-based systems into educational analysis offers a promising avenue for evaluating learning concentration in a more efficient and technologically driven framework.

Despite the well-established effectiveness of CNN models in image classification tasks, their overall performance remains highly contingent upon both the quality and the quantity of the training dataset. When the available data is limited, models are more prone to experiencing overfitting, a condition in which the system becomes excessively tailored to the training data and fails to perform adequately when exposed to new or unseen inputs. This limitation significantly reduces the model's generalization capability, thereby undermining its reliability in practical applications. Consequently, ensuring sufficient data variability and representation becomes a fundamental requirement in developing robust deep learning models for facial expression recognition tasks.

A widely adopted strategy to mitigate the challenges associated with limited datasets is the application of data augmentation techniques, which involve generating diverse variations of existing data to artificially expand the dataset. Through transformations such as rotation, scaling, or flipping, augmentation aims to enrich the training process by exposing the model to a broader range of patterns. However, the degree of complexity applied in these augmentation techniques plays a critical role in determining the effectiveness of the learning process. Excessive or inappropriate transformations may distort essential features, potentially leading to suboptimal model performance. Therefore, this study is designed to investigate how varying levels of data augmentation influence the performance of the MobileNetV2 model in the context of classifying students' facial expressions.

Literature Review

Data augmentation has become a fundamental approach in deep learning for enhancing model generalization by artificially expanding the variability of training datasets. As noted by Goodfellow et al. [1], achieving optimal performance in deep learning models generally requires access to large-scale data. In practice, however, many applications including facial expression recognition are constrained by datasets that are both limited in size and unevenly distributed across classes. To overcome these constraints, a range of augmentation techniques is frequently employed, such as rotation, horizontal flipping, scaling transformations, and adjustments in image brightness, all of which are intended to produce new data variations derived from existing samples [3]. Through this process, the model is exposed to a broader spectrum of input patterns, thereby improving its learning capacity without requiring additional data collection.

A number of prior investigations have confirmed that augmentation techniques can play a substantial role in strengthening model robustness while simultaneously mitigating overfitting. Perez and Wang [3], for instance, demonstrated that relatively simple augmentation strategies are capable of improving classification outcomes across multiple image-based datasets. Nevertheless, their findings also underscore that the success of augmentation is not uniform; rather, it is strongly influenced by both the type and the magnitude of the transformations applied. When augmentation is performed excessively or without careful consideration, it may introduce artificial or unrealistic patterns that deviate from the original

data distribution. Such distortions can interfere with the model's ability to learn meaningful representations, ultimately reducing its effectiveness.

Within the domain of facial expression recognition, the application of augmentation presents additional challenges due to the inherently subtle distinctions that exist between different emotional categories. According to Li and Deng [11], augmentation can contribute positively to recognition accuracy, but only when implemented in a manner that preserves critical facial characteristics. If essential features are altered or obscured, the model may struggle to differentiate between closely related expressions. Similarly, Wang et al. [7] highlight that high inter-class similarity and class imbalance remain persistent issues in commonly used datasets such as FER-2013. These factors further complicate the learning process and necessitate a more deliberate and controlled application of augmentation techniques.

MobileNetV2 represents a convolutional neural network architecture specifically engineered to balance computational efficiency with strong predictive performance [4]. Its design incorporates inverted residual structures alongside linear bottleneck layers, enabling it to operate effectively in environments with limited computational resources. Empirical evidence from previous studies indicates that MobileNet-based architectures perform competitively in a range of image classification tasks, including those involving facial expression analysis [12]. Despite these advancements, much of the existing research has concentrated primarily on architectural refinements, while comparatively little attention has been directed toward systematically evaluating the role of data augmentation within such models.

Although data augmentation is widely adopted as a standard preprocessing technique, there remains a notable gap in the literature regarding its nuanced impact on model performance, particularly when different levels of augmentation complexity are considered. Most prior works tend to implement augmentation as a routine step without examining how varying degrees of transformation intensity may influence learning outcomes or potentially introduce adverse effects. This lack of detailed analysis limits the understanding of how augmentation strategies can be optimized for specific model architectures and datasets.

In response to this gap, the present study seeks to conduct a structured examination of how different augmentation levels namely no augmentation, light augmentation, and complex augmentation affect the performance of the MobileNetV2 model in facial expression recognition tasks. By comparing these distinct scenarios, the study aims to identify patterns and relationships that clarify the role of augmentation intensity in shaping model behavior. The findings are expected to offer more precise guidance in determining augmentation strategies that enhance performance without compromising the integrity of learned features.

Recent contributions from local researchers further illustrate the expanding utilization of machine learning and deep learning techniques across diverse application domains. For example, Novelan and Aryza [13] investigated machine learning approaches for addressing imbalanced classification problems, a challenge that is directly relevant to datasets like FER-2013. In addition, Idhami et al. [14] demonstrated the effectiveness of neural network models in predictive tasks, highlighting their capability to capture complex and non-linear data relationships. These studies collectively emphasize the versatility and growing importance of deep learning methodologies in handling real-world data challenges.

Moreover, the practical implementation of deep learning-based facial recognition systems has been demonstrated in applied settings, such as attendance monitoring systems [15], reinforcing the real-world applicability of facial classification technologies. Complementary research by Yuliansyah et al. [16] further underscores the contribution of advanced artificial intelligence models in enhancing decision-making processes across various sectors. These applied perspectives are supported by foundational theoretical insights in machine learning as discussed by Novelan and Aryza [17], which provide a conceptual basis for understanding model behavior and optimization.

Taken together, these studies highlight the critical importance of carefully selecting both model configurations and data processing strategies, including augmentation techniques, in order to achieve improved classification performance. The integration of appropriate augmentation methods, when aligned with model architecture and dataset characteristics, plays a decisive role in ensuring robust and reliable deep learning outcomes.

Research Methodology

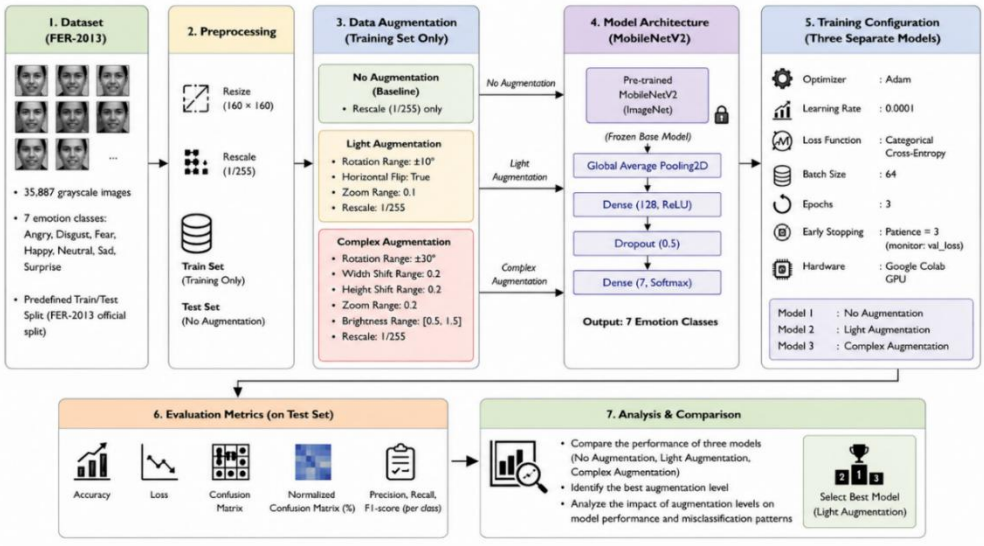

Figure 1. Research Workflow of Facial Expression Recognition

This study employs a deep learning approach to analyze the effect of data augmentation levels on facial expression classification performance. The overall workflow of the proposed method is illustrated in Fig. 1.

a. Dataset Description

The dataset used in this study is the FER-2013 dataset, which is widely utilized for facial expression recognition tasks. The dataset consists of grayscale facial images categorized into seven emotion classes: angry, disgust, fear, happy, neutral, sad, and surprise.

The dataset is divided into training and testing sets to evaluate the generalization capability of the model. Each class exhibits an imbalanced distribution, which presents a challenge for classification performance, particularly for minority classes such as *disgust*.

b. Data Preprocessing

All input images are resized to a fixed resolution of 160 × 160 pixels to ensure uniformity in model input dimensions. Furthermore, pixel values are normalized to a range of 0 to 1 using rescaling, which helps accelerate convergence during training and stabilizes the learning process.

c. Data Augmentation Strategy

To address the issue of limited data and improve model generalization, two levels of data augmentation are applied:

1. **Light Augmentation**

   This strategy includes mild transformations such as:

   a. Small rotations (±10°)
   b. Horizontal flipping
   c. Slight zooming (0.1)

These transformations preserve essential facial features while introducing moderate variability.

2. **Complex Augmentation**

This strategy applies more aggressive transformations, including:

a. Larger rotations (up to 30°)
b. Width and height shifting (0.2)
c. Zooming (0.2)
d. Brightness adjustment (0.5–1.5)

While complex augmentation increases data diversity, it may distort critical facial features, potentially affecting model performance.

It is important to note that augmentation is applied only to the training dataset, while the testing dataset remains unchanged to ensure a fair evaluation.

d. Model Architecture

The model used in this study is MobileNetV2, a lightweight convolutional neural network architecture designed for efficient computation and suitable for resource-constrained environments.

The base model is initialized using pre-trained weights from ImageNet and is frozen during training to retain learned feature representations. On top of the base model, several additional layers are added:

1. Global Average Pooling layer
2. Fully connected (Dense) layer with 128 neurons and ReLU activation
3. Dropout layer with a rate of 0.5 to reduce overfitting
4. Output layer with Softmax activation for multi-class classification

This architecture enables efficient feature extraction while maintaining low computational complexity.

e. Training Configuration

The model is trained under three different scenarios:

1. Without augmentation
2. With light augmentation
3. With complex augmentation

All scenarios use the same training configuration to ensure a fair comparison. The model is compiled using the Adam optimizer with categorical cross-entropy as the loss function.

The training process is conducted with:

1. Batch size: 64
2. Number of epochs: 3
3. Early stopping with patience of 3 epochs based on validation loss

Early stopping is employed to prevent overfitting by restoring the best model weights when validation performance stops improving.

All experiments are performed using Google Colab with GPU acceleration to enhance computational efficiency.

f. Evaluation Metrics

To evaluate the performance of the classification model, several standard metrics are used, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's effectiveness in multi-class classification tasks.

Accuracy measures the overall correctness of the model and is defined as:

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

Precision evaluates the proportion of correctly predicted positive samples among all predicted positives:

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

Recall (sensitivity) measures the proportion of correctly predicted positive samples among all actual positives:

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

The F1-score represents the harmonic mean of precision and recall:

$$\textbf{\textit{F1-score}} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the elements of the confusion matrix.

In addition, confusion matrices and normalized confusion matrices are used to provide detailed insights into classification performance across different emotion classes.

Results

Model Training Results

The experimental results show differences in model performance across each augmentation

scenario.The experimental results show differences in model performance across each augmentation scenario. Models without augmentation tend to have lower accuracy compared to models that use augmentation.

Light augmentation provides a significant improvement in model performance because it adds data variation without altering the primary characteristics of the facial images. Meanwhile, complex augmentation does not always yield the best results because overly extreme transformations can reduce the clarity of important features in the images.

Additionally, the use of augmentation has been shown to help reduce overfitting, as evidenced by a smaller difference between training and validation accuracy.

The model training results for the three scenarios are shown in the following table:

Table 1. Model Training Results

Scenario	Accuracy	Loss
No Augmentation	0.108	1.990
Light Augmentation	0.227	1.866
Complex Augmentation	0.083	1.847

The table shows that light augmentation yields the highest accuracy compared to the other scenarios.

a. Accuracy and Loss Analysis

Based on the accuracy plot, the model with light data augmentation performed best, achieving the highest validation accuracy of 22.7%. The model without data augmentation showed an initial improvement at the start of the epoch but then declined, indicating mild overfitting. Meanwhile, complex data augmentation showed a consistent decline in accuracy, indicating that overly extreme data transformations actually hinder the model's learning process.

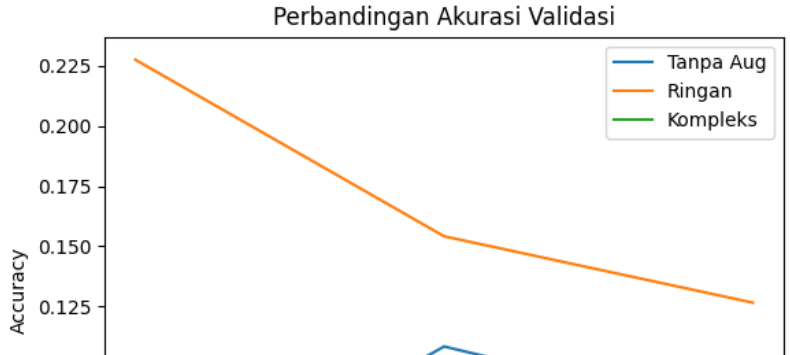

Figure 2. Validation Accuracy Graph
Figure 2. Validation Accuracy Graph

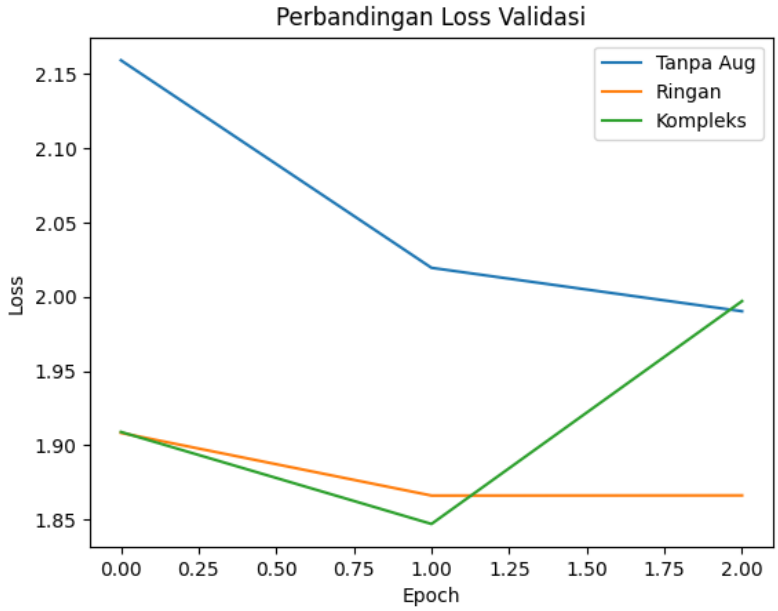

Figure 3. Validation Loss Graph

In terms of loss, complex augmentation yields the lowest loss value; however, this is not accompanied by an increase in accuracy, indicating that the model is unable to classify effectively despite the relatively small error.

b. Confusion Matrix Analysis

Table 2. Confusion Matrix (Light Augmentation)

Actual / Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	137	0	129	186	128	281	97
Disgust	11	0	18	26	14	33	9
Fear	110	0	190	145	103	294	182
Happy	126	0	122	950	145	303	128
Neutral	93	0	127	257	363	292	101
Sad	113	0	131	244	147	561	51

Surprise	44	0	128	61	55	52	491

The rows show the actual classes, while the columns show the model's predictions.

The confusion matrix obtained from the light augmentation model provides a detailed overview of the classification performance across seven emotion classes, as shown in Table 2.

The results indicate that the model performs relatively well in recognizing the *happy*, *sad*, and *surprise* classes, with correct predictions of 950, 561, and 491 samples, respectively. These classes exhibit more distinctive facial features, making them easier for the model to learn and generalize.

However, the model shows limited performance in distinguishing several other classes, particularly *disgust*, which is not predicted correctly at all. This suggests that the model fails to capture discriminative features for this class, likely due to insufficient or imbalanced training samples.

Furthermore, significant misclassifications are observed between visually similar expressions. For instance, *fear* is frequently misclassified as *sad* (294 samples) and *happy* (145 samples), while *neutral* is often confused with *happy* (257 samples) and *sad* (292 samples). These findings indicate that the model struggles to differentiate between subtle facial variations.

Overall, the confusion matrix reveals that although light augmentation improves general performance, the model still suffers from class imbalance and inter-class similarity, which significantly affect classification accuracy.

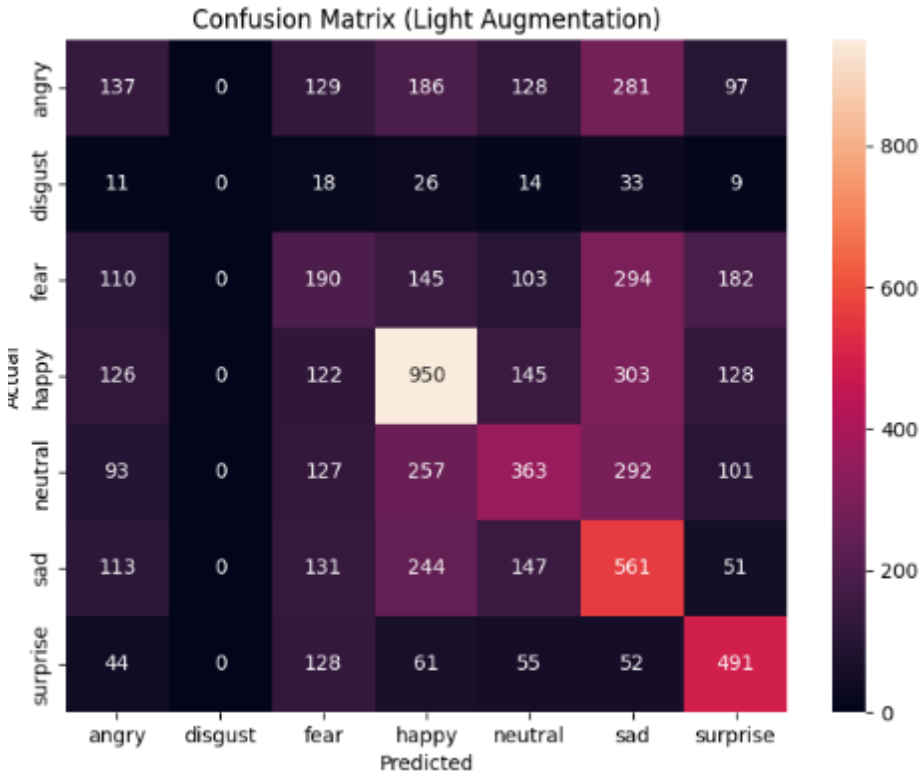

Figure 4. Confusion Matrix Heatmap for Light Augmentation Model

The confusion matrix illustrates the distribution of predicted labels across all facial expression classes for the light augmentation model. As shown in Fig. 4, the model achieves relatively higher correct predictions along the diagonal for the *happy*, *sad*, and *surprise* classes, indicating better recognition performance for these expressions. In particular, the *happy* class shows the highest number of correctly classified samples (950 instances), followed by *sad* (561 instances) and *surprise* (491 instances).

However, significant misclassifications are observed across several classes. For example, samples from the *angry*, *fear*, and *neutral* classes are frequently misclassified as *sad* and *happy*, indicating overlapping feature representations among these expressions. Additionally, the *disgust* class exhibits extremely poor recognition performance, with nearly all samples being misclassified into other categories.

These results suggest that the model tends to bias predictions toward dominant or visually distinguishable classes, while struggling to learn subtle facial variations in less represented or ambiguous categories. This imbalance highlights the limitations of the model when dealing with complex emotional expressions and reinforces the need for further optimization, such as class balancing or feature enhancement techniques.

c. Normalized Confusion Matrix Analysis

.

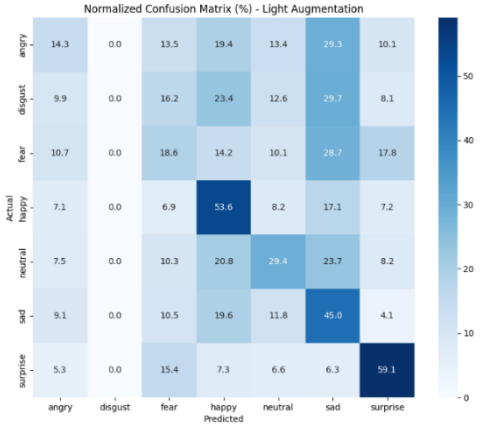

Figure 5. Normalized Confusion Matrix (%) for Light Augmentation Model

To provide a more interpretable evaluation, the confusion matrix is normalized into percentage values, representing the proportion of predictions for each class relative to the total number of actual samples. As shown in the normalized confusion matrix, the model demonstrates relatively strong performance in recognizing happy (53.6%), sad (45.0%), and surprise (59.1%) expressions. In contrast, performance on the disgust class is extremely poor (0%), indicating that the model fails to learn discriminative features for this category.

Furthermore, several classes exhibit significant misclassification patterns. For example, angry, fear, and disgust samples are frequently misclassified as sad (up to ~29%), suggesting that the model struggles to differentiate between negative facial expressions with similar visual characteristics. These results highlight the impact of class imbalance and the limitations of the model in capturing subtle inter-class variations. The normalized confusion matrix is presented for the best-performing model (light augmentation) to provide a clearer and more focused interpretation of classification behavior.

This observation suggests that further improvements such as class balancing, fine-tuning, or advanced architectures may be required to enhance model generalization.

d. Classification Report

Here is the evaluation table for each class:

Table 3. Classification Performance

Class	Precision	Recall	F1-Score
Angry	0.20	0.14	0.17
Disgust	0.00	0.00	0.00
Fear	0.18	0.17	0.17
Happy	0.53	0.54	0.53
Neutral	0.28	0.26	0.27

Sad	0.28	0.31	0.29
Surprise	0.44	0.50	0.47

Based on the evaluation results using precision, recall, and F1-score, it is evident that the Happy and Surprise classes achieved the best performance among all classes. Conversely, the Disgust class was not detected at all by the model, as indicated by precision and recall values of 0. This suggests data imbalance and the model's difficulty in recognizing features in certain classes.

e. Discussion

The results of the study show that light augmentation yields the best performance in improving model accuracy. This is because light augmentation increases data variability without altering the primary characteristics of the facial images.

In contrast, complex augmentation does not yield a significant improvement in performance. Transformations that are too extreme cause distortion in facial features, making it difficult for the model to recognize relevant patterns.

The results from the confusion matrices across all scenarios reveal a significant difference in model behavior. The model trained without augmentation tends to overfit and shows a strong bias toward dominant classes such as *happy*.

In contrast, the model with light augmentation demonstrates a more balanced distribution of predictions across classes, indicating improved generalization capability.

However, the model with complex augmentation exhibits a severe bias toward the *sad* class, suggesting that excessive transformations distort important facial features and negatively impact learning. This phenomenon indicates that overly aggressive augmentation can lead to model degradation rather than performance improvement.

Additionally, the relatively low accuracy scores across all scenarios indicate that the model is not yet optimized. This is likely due to several factors, including:

1. A limited number of training epochs
2. No fine-tuning was performed on the MobileNetV2 base model
3. Imbalance in the amount of data across each class

The findings of this study are consistent with previous research showing that data augmentation can improve model performance [3]. However, unlike some studies that suggest more complex augmentation leads to better results, this study demonstrates that excessive augmentation may degrade performance. This result aligns with the findings of Perez and Wang [3], which state that overly aggressive transformations can distort important image features. Therefore, selecting an appropriate augmentation strategy is crucial for achieving optimal performance.

The relatively low accuracy suggests that further optimization is necessary, particularly through fine-tuning and dataset balancing techniques.

f. Recommendations for Further Development

To improve model performance in future research, the following steps can be taken:

1. Increase the number of training epochs
2. Perform fine-tuning on the MobileNetV2 layers
3. Use data balancing techniques such as oversampling
4. Try other model architectures such as ResNet or EfficientNet

Conclusion

This study confirms that the degree of data augmentation plays a crucial role in determining the performance of the MobileNetV2 model for facial expression classification tasks. The results indicate that light augmentation yields the most favorable outcomes, suggesting that moderate and controlled transformations are capable of enriching data variability while still preserving the essential features required for accurate classification. By maintaining the integrity of key visual characteristics, this level of augmentation allows the

model to learn more effectively without introducing unnecessary noise or distortion. Consequently, light augmentation emerges as a balanced approach that supports both improved generalization and stable learning performance.

In contrast, the application of complex augmentation techniques is shown to have a detrimental effect on model performance. The excessive and aggressive transformations involved in this approach tend to distort important image attributes, making it more difficult for the model to extract meaningful patterns from the data. As a result, the model's ability to classify facial expressions accurately is significantly reduced. These findings emphasize that augmentation must be applied with careful consideration, as inappropriate strategies can lead to unintended consequences that undermine model effectiveness. Overall, the study underscores the necessity of selecting augmentation methods that are well-aligned with the characteristics of the dataset and the objectives of the model.

Looking ahead, future research can explore several avenues to further enhance model performance, including fine-tuning model parameters, extending the number of training epochs, and implementing class balancing techniques to address dataset imbalances. In addition, the experimental outcomes reinforce the notion that poorly designed augmentation strategies can adversely affect learning results, highlighting the importance of systematic evaluation when integrating augmentation into deep learning workflows.

References

[1] Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[2] Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. NeurIPS, 2012.

[1] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification," arXiv preprint arXiv:1712.04621, 2017.

[2] M. Sandler et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," CVPR, 2018.

[3] Goodfellow et al., "Challenges in Representation Learning: A Report on FER-2013," in ICML Workshop, 2013.

[4] S. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, 2020.

[5] H. Wang, Y. Wang, and Z. Zhou, "Facial Expression Recognition Based on Deep Learning: A Survey," IEEE Access, vol. 8, pp. 146, 2020.

[6] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, 2020.

[7] Howard et al., "Searching for MobileNetV3," in Proc. ICCV, 2020.

[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial Expression Recognition Using Deep Learning: A Review," Pattern Recognition, vol. 122, 2022.

[9] J. Li and S. Deng, "Robust Facial Expression Recognition with Data Augmentation and CNN," IEEE Access, vol. 10, pp. 2023.

[10] Y. Li et al., "Lightweight CNN-Based Facial Expression Recognition for Real-Time Applications," Sensors, vol. 23, no. 4, 2023.

[11] M. S. Novelan and S. Aryza, "Optimization CVRP with Machine Learning for Improved Classification of Imbalanced Data Food Distribution," JITK, vol. 10, no. 4, pp. 917–925, 2025.

[12] R. Idhami et al., "Implementation of an Intelligent System to Predict Product Demand with Backpropagation Neural Network Algorithm," IJSET, 2026.

[13] N. Tampubolon et al., "Implementation of Deep Learning Algorithm for Face Detection Attendance System," IJSET.

[14] R. Yuliansyah et al., "Deep Q-Network Analysis in Optimizing Data Processing for Decision Making," Jurnal Jatilima, 2025.

[15] M. S. Novelan and S. Aryza, “Belajar Mengenal Dasar Machine Learning”, Serasi Media Teknologi, 2025.