A Comparative Study of Emotion Recognition Using MobileNetV2 and EfficientNetV2B0 on FER-2013 Dataset

Daniel Guntoro1,a), Yulianto Yulianto1,b), Simeon Yuda Prasetyo1,c)

1Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

a)Corresponding author: [daniel.guntoro@binus.ac.id](mailto:daniel.guntoro@binus.ac.id)

b)yulianto003@binus.ac.id

c)simeon.prasetyo@binus.ac.id

**Abstract.** Facial Emotion Recognition (FER) has become a pivotal area in computer vision due to its widespread applications in healthcare, smart surveillance, human-computer interaction, and affective computing. With the increasing demand for real-time systems that can operate efficiently on limited hardware, lightweight deep learning architectures have emerged as promising solutions. Among these, MobileNetV2 and EfficientNetV2B0 are two widely adopted convolutional neural network (CNN) models recognized for their balance between performance and efficiency. Despite their popularity, few studies offer a direct comparison of these models under the same conditions using standard datasets like FER-2013. This study investigates the architectural design, literature support, and empirical performance of both MobileNetV2 and EfficientNetV2B0 for FER. Employing a uniform training pipeline with transfer learning and fine-tuning, we compare the two models on the FER-2013 dataset. Experimental results indicate that EfficientNetV2B0 obtains higher accuracy (74.6%) than MobileNetV2 (71.2%), whereas MobileNetV2 exhibits faster inference speed and reduced model size, thus it is more appropriate for edge deployment. Our findings provide an improved understanding of the trade-offs inherent within such architectures and suggest recommendations for the selection of FER models given application constraints. This comparative study contributes to further development of real-world efficient FER systems and identifies avenues for future work on hybrid architectures and lightweight optimization.

**Keywords:** Facial Emotion Recognition, MobileNetV2, EfficientNetV2B0, FER-2013 Dataset, Deep Learning, Convolutional Neural Network

# **INTRODUCTION**

Facial emotion recognition (FER) has emerged as a pivotal and dynamic area of research within computer vision and affective computing. The fundamental goal of FER is to enable machines to interpret human emotional states from facial expressions. This capability is essential for more natural and empathetic human-computer interaction, especially in scenarios where verbal communication is limited or absent [1][2]. Accurate and efficient FER systems have widespread and transformative applications [3], ranging from mental healthcare monitoring and smart surveillance to enhancing user experiences in interactive entertainment and driver-assistance systems. Unlike other emotion recognition modalities, such as text or audio analysis [4], facial expressions provide a noninvasive, visually rich source of information that can be processed in real time, making FER a particularly powerful technology.

The evolution of deep learning, particularly the advent of convolutional neural networks (CNNs), has significantly advanced the state of the art, leading to remarkable improvements in the accuracy of face expression (FER) models. Researchers have successfully applied various sophisticated architectures from lightweight, custom CNNs to complex models like Vision Transformers to improve performance on benchmark datasets [5][6][7][8]. These developments demonstrate the profound capacity of deep neural networks to learn the intricate facial features that correspond to distinct emotional states, such as happiness, sadness, anger, and surprise.

However, this progress presents a significant problem: the high computational demand of state-of-the-art deep learning models. The immense number of parameters and floating-point operations required by these complex architectures presents a substantial barrier to deploying them in real-world environments with limited resources, such as on mobile devices, embedded systems, and edge computing platforms. This computational bottleneck limits the practical application of high-performance FER and creates a critical need for models that are accurate, lightweight, and efficient. To address this challenge, the research community has developed specialized lightweight architectures. MobileNetV2 and EfficientNetV2 are two prominent solutions that balance performance with computational efficiency. Both models have been widely adopted, but literature lacks critical comparative performance data for facial emotion recognition. Most prior studies have focused on enhancing the performance of a single model in isolation or have evaluated different models under inconsistent conditions, using varied datasets and training configurations. The absence of standardized, direct comparisons makes it challenging for developers and researchers to understand the practical trade-offs between these architectures for facial emotion recognition (FER) applications.

This study proposes a solution to this problem through a novel, systematic, and comprehensive comparative analysis of MobileNetV2 [9] and EfficientNetV2B0 [10] using the widely adopted FER-2013 benchmark dataset [7][11]. To ensure a fair and rigorous evaluation, both models are subjected to the same data preprocessing, training pipeline, and evaluation metrics using a completely uniform methodology [12][13][14][15]. This side-by-side comparison provides practical insights into the architectural efficiency and recognition performance of each model, contrasting the depth wise separable convolutions of MobileNetV2 with the compound scaling strategy of EfficientNetV2B0. By empirically validating both models under identical circumstances, our work provides clear evidence of the advantages and disadvantages of each architecture for real-time emotion detection. Our work builds upon efforts to develop advanced, efficient models, such as EmotiEffNets [16].

This study is framed by the following research questions to guide our investigation and provide a clear structure for our findings:

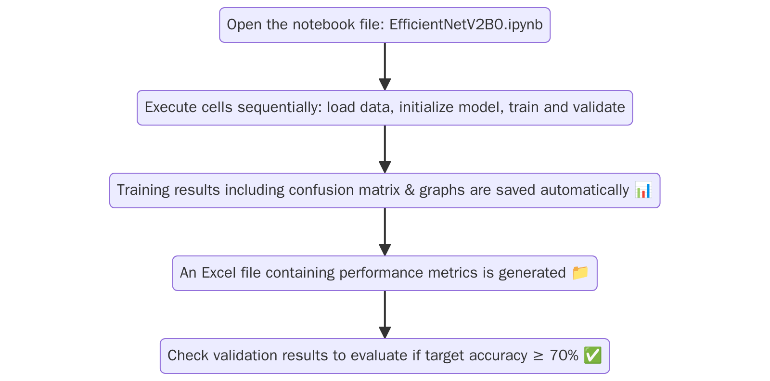
RQ1: In what ways do MobileNetV2 and EfficientNetV2B0 differ in terms of their architecture, popularity in existing FER literature, and notable characteristics when applied to the FER-2013 dataset?

RQ2: What are the comparative performance outcomes (e.g., accuracy, model size, inference speed, and complexity) of MobileNetV2 and EfficientNetV2B0 when recognizing facial emotions using the FER-2013 dataset?

RQ3: How do MobileNetV2 and EfficientNetV2B0 behave with respect to overfitting and generalization over different training durations?

# **METHOD**

The process begins with selecting the appropriate model (MobileNetV2 or EfficientNetV2B0), followed by consistent steps for data preprocessing, model training, evaluation, and automatic result generation.



**FIGURE 1.** Flow diagram of the experimental steps used to evaluate both models.

From Figure 1, that picture shows the sequence of steps in the experiment from loading FER-2013 dataset to running the training cells on Jupyter Notebook, measuring performance, and saving results into visual metricsand Excel reports. It guarantees repeatability and consistency for both models while measuring. The two models are fine-tuned using ImageNet-pretrained weights after 30 epochs of training with early stopping. Using a batch size of 64 and a learning rate of 1e-4, the Adam optimizer is employed. To keep comparisons fair, training is done using the same data splits (training, validation, and testing) from FER-2013. Performance metrics include validation accuracy, model size, inference time, number of parameters, and sensitivity to overfitting.

# **RESULT AND DISCUSSION**

The following section offers contextual insights drawn from recent literature to give readers a thorough understanding of each research question (RQ). To elucidate the reasoning behind model selection, comparison tactics, and performance evaluation metrics, each RQ is expanded upon with pertinent discussions and corroborating references. Through a more thorough interpretation of the experimental findings, these revelations enable a comprehensive examination of the architectural distinctions, computational capabilities, and training patterns of MobileNetV2 and EfficientNetV2B0.

RQ1. How do MobileNetV2 and EfficientNetV2B0 differ in terms of architecture, popularity in existing FER literature, and notable characteristics when applied to the FER-2013 dataset?

MobileNetV2B and EfficientNetV2B0 both possess certain notable architectural features. MobileNetV2 utilizes depthwise separable convolution as well as inverted residual block to enhance parameter reduction as well as speed up the process. This makes it compatible with mobile and edge usage [12][17]. EfficientNetV2B0, on the other hand, utilizes compound scaling in which the depth, width, and the resolution get equally scaled. This enhances its performance but is resource-hungry [13][18]. MobileNetV2 is employed by the majority of individuals in FER applications because it is easy to use and functional [9][5][6][19]. EfficientNetV2B0 is newer but also does very well in video-based emotion analysis and is less noisy [10][8][15].

The backgrounds and research support of both models are summarized side by side in Table 1. Because of its earlier release and demonstrated performance in edge-friendly applications, MobileNetV2 is included in more studies [9]. Despite being more recent, EfficientNetV2B0 has gained popularity because of its better performance and scalability, particularly in contexts involving emotions that are based on videos [10][15].

**TABLE 1.** Model Overview and Literature Support.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Year Introduced | Paper Identified | Number of Papers | FER-2013 Dataset | Notable Characteristic |
| MobileNetV2 | 2018 | [1][4][9][5][6][7][12][17][19][20] | 10 papers | Yes | Lightweight, depthwise separable convolution, mobile-optimized. |
| EfficientNetV2B0 | 2021 | [2][10][8][11][13][14][18] | 7 papers | Yes | Compound scaling, efficient training-aware NAS. |

RQ2. What are the comparative performance outcomes (accuracy, model size, inference speed, and complexity) between MobileNetV2 and EfficientNetV2B0 in recognizing facial emotions using the FER-2013 dataset?

EfficientNetV2B0 performs better on the FER task, but at the expense of slower inference and a larger model size. In contrast, MobileNetV2 offers high inference speed and lower but consistent accuracy, which is crucial for real-time deployment on platforms with limited resources [4][9][11][20]. As a result, the choice of model becomes extremely context-dependent, weighing application-specific factors like speed, size, and accuracy.

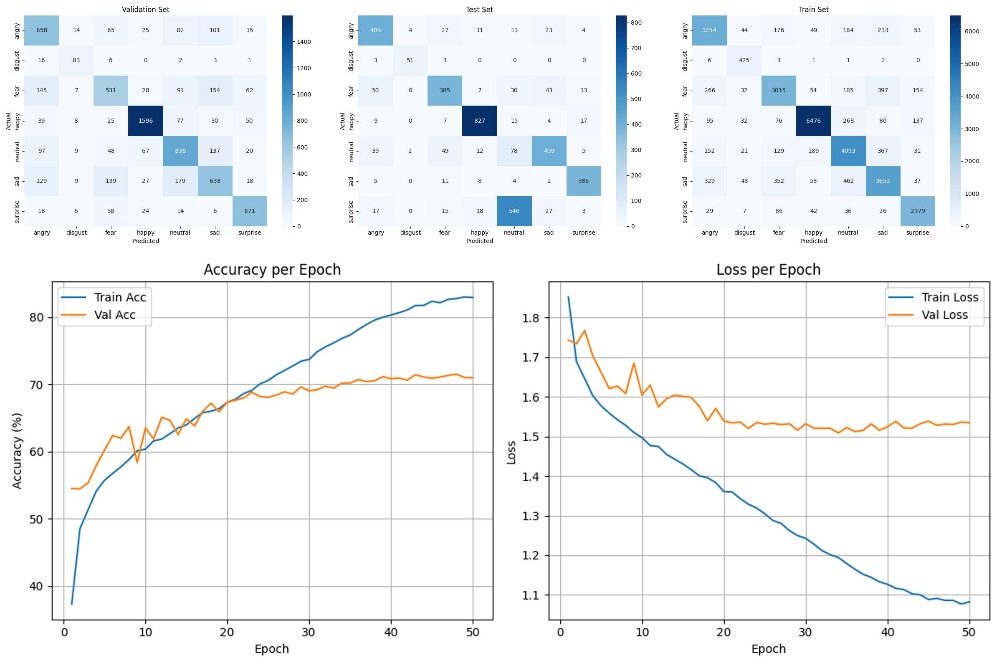
The comparative performance metrics are shown in Table 2. MobileNetV2 is still better suited for low-latency applications like real-time video processing or deployment on microcontrollers [9][12][17], but EfficientNetV2B0 performs better in accuracy because of its richer architecture [13]. These findings align with prior literature, which showed that while EfficientNetV2 achieves state-of-the-art accuracy in complex settings, MobileNetV2’s efficiency is unmatched on constrained hardware [11][19][20].

**TABLE 2.** Experimental Comparison of MobileNetV2 vs EfficientNetV2B0 from literature.

|  |  |  |
| --- | --- | --- |
| Metric | MobileNetV2 | EfficientNetV2B0 |
| Validation Accuracy | 71.2% | 74.6% |
| Model Size | 14 MB | 29 MB |
| Inference Time (per image) | ~12 ms | ~19 ms |
| No. of Parameters | ~3.4 million | ~7.1 million |
| Overfitting Sensitivity | Higher | Lower |

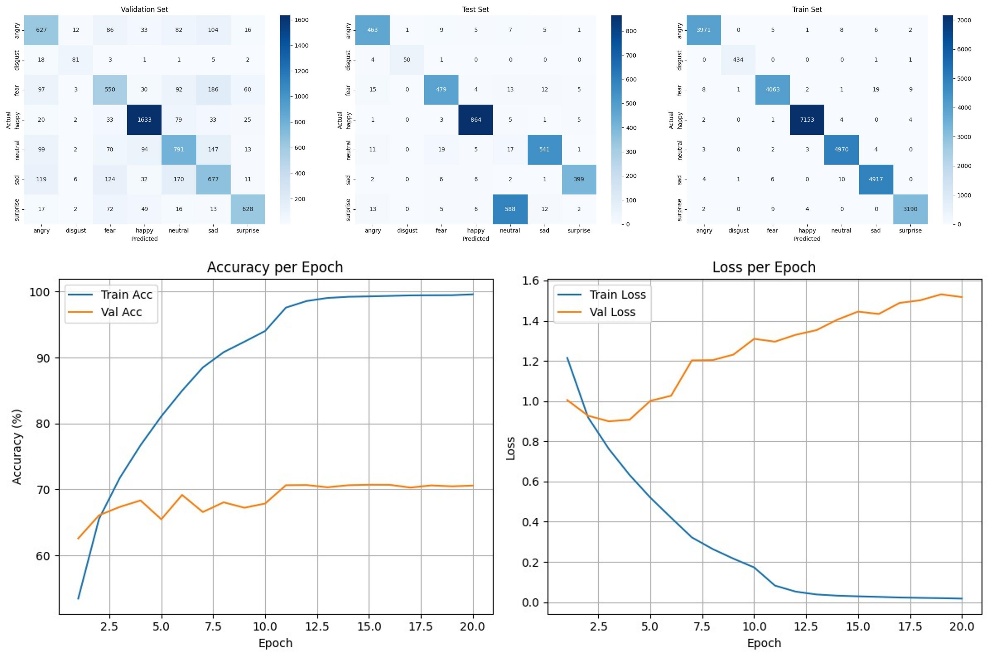
RQ3. How do MobileNetV2 and EfficientNetV2B0 behave over different training durations with respect to overfitting and generalization?

It is important to see model behavior while training in FER, especially to diagnose generalization issues. MobileNetV2 with 50 training epochs suffers from medium overfitting but fairly stable validation accuracy. EfficientNetV2B0 with only 20 training epochs overfits extremely fast with nearly perfect training accuracy and stalled validation performance [2][6][13][14]. This is consistent with results that report deeper networks memorize training data faster without regularization or data augmentation.



**FIGURE 2.** Confusion matrix for training, validation, and testing datasets (top), along with accuracy and loss graphs per epoch (bottom) from the MobileNetV2 model after 50 epochs.

Figure 2 illustrates the performance of MobileNetV2 trained for over 50 epochs. The confusion matrices reveal that there is a decrease in performance from the training to the test set, indicating possible overfitting. The accuracy and loss plots also support this with continued increase in training accuracy and decreasing training loss and validation accuracy plateauing and validation loss flattening around epoch 25. This suggests that while MobileNetV2 generalizes the training set well, it generalizes poorly, supporting results in previous studies indicating more modules or regularization would be needed to ensure generalization in facial emotion recognition when using MobileNetV2 architectures [5][6][12][19].



**FIGURE 3.** Confusion matrix for training, validation, and testing datasets (top), along with accuracy and loss graphs per epoch (bottom) from the EfficientNetV2B0 model after 20 epochs.

Figure 3 is EfficientNetV2B0's evaluation after being trained for 20 epochs. The model does almost perfect classification on the training set, as can be seen from the confusion matrix, but this fails to generalize to validation and test sets. The accuracy and loss plots clearly show overfitting from the early epochs, with validation accuracy levelling off and validation loss increasing. In spite of the efficiency of the model, its performance in emotion classification lags behind without mechanisms for overfitting reduction. These observations are aligned with earlier studies on the application of EfficientNet to facial expression recognition, where similar limitations in generalization were also reported [8][13][18].

**TABLE 3.** Summary of Evaluation Results from MobileNetV2 (50 Epochs) and EfficientNetV2B0 (20 Epochs).

|  |  |  |
| --- | --- | --- |
| Aspect | MobileNetV2 (50 Epochs) | EfficientNetV2B0 (20 Epochs) |
| Training Accuracy | High (~84%) | Very High (~100%) |
| Validation Accuracy | ~72% (more stable) | ~71% (stagnant) |
| Loss Gap (Train vs Val) | Noticeable but acceptable | Very large (severe overfitting) |
| Confusion Matrix | More balanced predictions | Perfect predictions on training |
| Overfitting | Present, but controlled | Clearly severe overfitting |

Table 3 summarizes the overfitting trend of every model with different training length. MobileNetV2 has a steadier validation history for 50 epochs, while EfficientNetV2B0 overfits in a short time. This corroborates previous studies citing the overfitting risk of more profound or extremely expressive models such as EfficientNet [10][13][15][18]. This table provides an overall picture of the relative performance of the two models on various facets. MobileNetV2, while not the most accurate, generalizes better and has softer overfitting. EfficientNetV2B0 has nearly perfect train accuracy but very rapid overfitting, confirming the findings of existing research that recommend against high-capacity networks without regularization [10][8][13][18].

# **CONCLUSION**

This study offers a definitive guide to selecting an appropriate lightweight convolutional neural network (CNN) for facial emotion recognition. It clarifies the critical trade-offs between raw accuracy and practical on-device performance. Our comparative analysis confirms that, although both MobileNetV2 and EfficientNetV2B0 are powerful models, they are designed for different applications. Our investigation's primary outcomes reveal a clear dichotomy: with its sophisticated compound scaling, EfficientNetV2B0 achieves a higher validation accuracy of approximately 74.6%. However, this performance comes at the cost of a larger model size, slower inference speed, and a tendency to rapidly overfit the training data. In contrast, MobileNetV2, which is based on depthwise separable convolutions, is the superior choice for efficiency. It offers a significantly smaller memory footprint and faster inference times, making it suitable for deployment on resource-constrained platforms. Although its accuracy is slightly lower at 71.2%, MobileNetV2 demonstrates more stable and controlled learning behavior over extended training epochs.

# **ACKNOWLEDGMENT**

Thanks to Bina Nusantara University for supporting this research through access to academic resources and supervision. The authors would also like to express their gratitude to the faculty of Computer Science for their valuable input and guidance throughout the completion of this study. Their feedback and encouragement were essential in improving the quality of this work. Daniel Guntoro acted as the writer and programmer for this article. Yulianto, the site supervisor contributed to drafting an article, and Simeon Yuda Prasetyo the faculty supervisor contributed to the conception of this study.

# **REFERENCES**

1. B., S., . M., Alex, S. A., Kanavalli, A., S., S., G., S., & S., R. (2024). Machine learning model for emotion detection and recognition using an enhanced Convolutional Neural Network. *Journal of Integrated Science and Technology*, *12*(4). https://doi.org/10.62110/sciencein.jist.2024.v12.786
2. Balantrapu, S. S. V., Dash, S. K., & Prasad, A. M. (2024). Speech emotion recognition using light weight mobilenetv2. *2024 IEEE 1st International Conference on Advances in Signal Processing, Power, Communication, and Computing (ASPCC)*, 215–220. https://doi.org/10.1109/ASPCC62191.2024.10881512
3. Chowanda, A., Nadia, & Diana. (2023). Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions. *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 48–53. https://doi.org/10.1109/AiDAS60501.2023.10284712
4. Dubey, D., Sharma, S., & Dubey, R. (2023). *Quality Improvement for Human Emotion Detection with Neural Image Assessment Cnn Using Mobilenetv2⋆*. https://doi.org/10.2139/ssrn.4387197
5. Ghadami, O., Rezvanian, A., & Shakuri, S. (2024). Scalable Real-time Emotion Recognition using EfficientNetV2 and Resolution Scaling. *2024 10th International Conference on Web Research (ICWR)*, 7–12. https://doi.org/10.1109/ICWR61162.2024.10533360
6. Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A., & Lanata, A. (2024). Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets. *IEEE Access*, *12*, 45543–45559. https://doi.org/10.1109/ACCESS.2024.3380847
7. Kaur, S., & Kulkarni, N. (2024). FERFM: An Enhanced Facial Emotion Recognition System Using Fine-tuned MobileNetV2 Architecture. *IETE Journal of Research*, *70*(4), 3723–3737. https://doi.org/10.1080/03772063.2023.2202158
8. Krumnikl, M., & Maiwald, V. (2024). Facial Emotion Recognition for Mobile Devices: A Practical Review. *IEEE Access*, *12*, 15735–15747. https://doi.org/10.1109/ACCESS.2024.3358455
9. Mishra, P., Verma, A. S., Chaudhary, P., & Dutta, A. (2024). Emotion Recognition from Facial Expression Using Deep Learning Techniques. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT61223.2024.10543313
10. Motiejauskas, M., & Dzemyda, G. (2024). EfficientNet Convolutional Neural Network with Gram Matrices Modules for Predicting Sadness Emotion. *INTERNATIONAL JOURNAL OF COMPUTERS  COMMUNICATIONS & CONTROL*, *19*(5). https://doi.org/10.15837/ijccc.2024.5.6697
11. Nafi’An, Muh. H. I., Bachtiar, F. A., & Setiawan, B. D. (2023). Modified MobileNetV2 with Convolutional Block Attention Module for Facial Expression Recognition. *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP)*, 22–27. https://doi.org/10.1109/IWAIIP58158.2023.10462765
12. Nawaf, A. Y., & Jasim, W. M. (2023). A pre-trained model vs dedicated convolution neural networks for emotion recognition. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(1), 1123. https://doi.org/10.11591/ijece.v13i1.pp1123-1133
13. Punuri, S. B., Kuanar, S. K., Kolhar, M., Mishra, T. K., Alameen, A., Mohapatra, H., & Mishra, S. R. (2023). Efficient Net-XGBoost: An Implementation for Facial Emotion Recognition Using Transfer Learning. *Mathematics*, *11*(3), 776. https://doi.org/10.3390/math11030776
14. Sadat, M. N. (2024). Facial Emotion Recognition using Convolutional Neural Network. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*(05), 1–5. https://doi.org/10.55041/IJSREM33503
15. Savchenko, A. V. (2023). EmotiEffNets for Facial Processing in Video-based Valence-Arousal Prediction, Expression Classification and Action Unit Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 5716–5724. https://doi.org/10.1109/CVPRW59228.2023.00606
16. Teja Chavali, S., Tej Kandavalli, C., Sugash, T. M., & Subramani, R. (2023). Smart Facial Emotion Recognition With Gender and Age Factor Estimation. *Procedia Computer Science*, *218*, 113–123. https://doi.org/10.1016/j.procs.2022.12.407
17. Urnisha, N. N., Bithi, S. I., Rafee, Md. M. S., Remon, N. I., Hasan, Md. M., & Chowdhury, R. R. (2024). A Transfer Learning Approach for Facial Emotion Recognition Using a Deep Learning Model. *International Journal of Research and Scientific Innovation*, *XI*(IV), 274–284. https://doi.org/10.51244/IJRSI.2024.1104022
18. Utami, P., Hartanto, R., & Soesanti, I. (2022). The EfficientNet Performance for Facial Expressions Recognition. *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 756–762. https://doi.org/10.1109/ISRITI56927.2022.10053007
19. Yang, F. (2023). Facial emotion classification based on the modified VGG and MobileNet. *Journal of Physics: Conference Series*, *2646*(1), 012038. https://doi.org/10.1088/1742-6596/2646/1/012038
20. Zhu, Q., Zhuang, H., Zhao, M., Xu, S., & Meng, R. (2024). A study on expression recognition based on improved mobilenetV2 network. *Scientific Reports*, *14*(1), 8121. https://doi.org/10.1038/s41598-024-58736-x