Fatigue Detection Based on Facial Expressions Using ResNet50 with Grad-CAM Visualization

Wike Septiana Vidya Utami [1], Ervin Yohannes [2],

[1,2] Teknik Informatika, Universitas Negeri Surabaya
[1]wike.22106@mhs.unesa.ac.id
[2]ervinyohannes@unesa.ac.id

Abstract*—Fatigue is a condition that can affect a person's performance and concentration, making an automatic system necessary for its rapid and accurate detection. This study applies a Convolutional Neural Network (CNN) architecture based on the pretrained ResNet50 model to classify facial images into two categories, fatigued and non-fatigued, using the Driver Drowsiness Dataset, which consists of approximately 41,700 images. The model was trained and validated using five data-split scenarios, namely 50:50, 60:40, 70:30, 80:20, and 90:10, with varying learning rates, batch sizes, and numbers of epochs. Performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results show that the best performance of the ResNet50 model was obtained with a 60:40 data split, a learning rate of 0.0001, a batch size of 16, and 30 epochs, achieving an accuracy of 99.22%, a precision of 99.67%, a recall of 99.86%, and an F1-score of 99.26%. Visual analysis using Grad-CAM showed that the model focused attention on the eyes, eyelids, cheeks, and forehead during the classification decision, improving the interpretability of the predictions. The ResNet50 model was also implemented in a camera-based system to provide real-time fatigue predictions. These results indicate that ResNet50 is effective and has the potential to be applied in a practical and efficient facial-image-based fatigue detection system.

Keywords*—ResNet50, Convolutional Neural Network (CNN), Fatigue Detection, Grad-CAM, Facial Image Classification.

I. INTRODUCTION

Fatigue is a condition of decreased physical or cognitive capacity that can affect a person's concentration and work performance. Several studies have shown a correlation between fatigue levels and an increased risk of occupational accidents and reduced productivity [1]. indicating the need for an automatic, fast, and accurate fatigue detection system.

Conventional image processing methods generally rely on manual feature extraction, which is sensitive to variations in lighting, camera angle, and individual characteristics [2]. Convolutional Neural Networks (CNN) offer a solution to these limitations, as they can automatically and hierarchically extract features directly from raw images, making them more adaptive in recognizing the complex facial expression patterns associated with fatigue [3].

One of the CNN architectures widely used in facial expression detection is the Residual Network (ResNet50). ResNet50 introduces the concept of residual learning with skip connections that allow very deep networks to be trained without suffering from performance degradation [4]. Skip connections preserve the flow of signals from previous layers while still allowing new layers to learn additional feature transformations, thereby improving the model's accuracy and generalization ability [5]. A study on facial expression recognition using ResNet-50 demonstrated that this architecture is effective as a feature extractor and can achieve high accuracy in facial emotion recognition [4].

In addition to classification performance, the interpretability of deep learning models is also an important concern. Grad-CAM (Gradient-weighted Class Activation Mapping) plays a role in explaining a CNN model's decisions by highlighting the image regions that most influence the detection result [6], making the predictions easier to understand and trust.

Based on the above, this study aims to build and evaluate a facial-expression-based fatigue detection model using the ResNet50 architecture equipped with Grad-CAM visualization, and to examine the model's performance across various training and validation data-split scenarios. The discussion in this article is specifically focused on the ResNet50 architecture in order to provide an in-depth study of its characteristics and performance in the task of facial-image-based fatigue detection

II. LITERATUR RIVIEW

A. *Convolutional Neural Networks (CNN)*

Convolutional Neural Networks (CNNs) are a deep learning architecture widely adopted for image analysis and have become a fundamental approach in computer vision tasks, including object detection, facial recognition, and image classification [7]. CNNs operate through two main stages: feature extraction and classification. During feature extraction, convolutional, activation [8], and pooling layers automatically learn discriminative visual features from input images. These extracted features are then transformed into a one-dimensional vector and passed through fully connected and softmax layers to predict the final image class [9].

B. *ResNet50*

Residual Network (ResNet) is a deep learning architecture designed to facilitate the training of very deep neural networks by addressing the vanishing gradient problem. It introduces residual blocks with skip connections [10] as shown in Figure 1

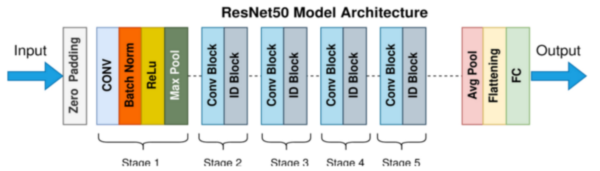

Figure 1. Arsitektur ResNet50

which enable information to bypass one or more layers, improving gradient flow and allowing deeper networks to achieve high performance without compromising accuracy [5].

C. *Grad-Cam*

Gradient-weighted Class Activation Mapping (Grad-CAM) is a visualization technique that explains CNN predictions by highlighting the image regions that contribute most to a target class [11]. It generates a class-specific activation map by computing the gradients of the target class with respect to the feature maps of the final convolutional layer and applying global average pooling to obtain the importance weights for each feature map [12].

D. *Fatigue*

Fatigue is a physiological and psychological condition characterized by a reduced ability to maintain alertness, cognitive performance, and motor function due to prolonged exertion or insufficient recovery time [13].

III. Research Method

A. *Dataset*

This study uses the Driver Drowsiness Dataset sourced from Kaggle The dataset represents two main conditions, namely fatigued (drowsy) and non-fatigued (non-drowsy), based on the driver's facial expression. The total data consists of approximately 41,700 images, comprising around 22,300 images in the fatigued class and 19,400 images in the non-fatigued class. Each image contains variations in lighting conditions, capture angles, facial expressions, and backgrounds, helping the model learn the visual patterns of fatigue more comprehensively

B. *Data Splitting and Preprocessing*

The dataset was divided into a training set and a validation set using five proportion scenarios, namely 90:10, 80:20, 70:30, 60:40, and 50:50, to analyze the effect of data proportion on the model's generalization ability. In the preprocessing stage, all images were resized to 224 × 224 pixels in accordance with ResNet50's input specification, converted into tensors, and then normalized using ImageNet parameters so that the feature-extraction process could run optimally.

C. *Training*

The design flow of the ResNet50 model in this study includes dataset import, splitting into training and validation data, image preprocessing, DataLoader construction, ResNet50 model building, configuration of the Adam optimizer and CrossEntropyLoss function, training and validation of the model at each epoch, saving the best model, performance evaluation, and finally testing predictions on new images, as illustrated in Figure. 2.

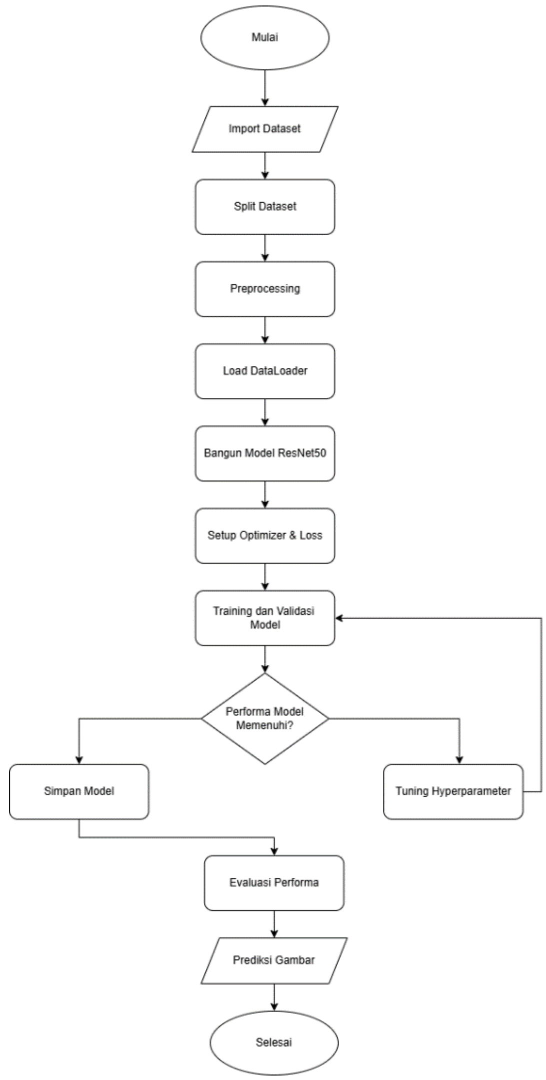

Figure 2. ResNet50 Model Design Flow

The parameter configuration used in training the ResNet50 model is presented in Table I. If the model's performance on the validation stage does not meet the required criteria, hyperparameter tuning is performed by adjusting the learning rate, batch size, and number of epochs until an optimal performance is obtained.

TABLE I

RESNET50 MODEL CONFIGURATION

Parameter	Specification / Value	Description
Model Architecture	ResNet50	CNN model pretrained on ImageNet
Input Size	224 × 224 pixels	Standard input size for the ResNet50 architecture
Number of Classes	2 classes	Fatigued and non-fatigued

Batch Size	8 – 32	Adjusted to GPU memory capacity
Learning Rate	1×10^{-3}, 1×10^{-4}	Small learning-rate values to maintain training stability
Number of Epochs	10 – 30 epochs	Maximum limit of the training process
Data Split	50:50, 60:40, 70:30, 80:20, 90:10	Used to evaluate the model's generalization ability
Optimizer / Loss	Adam / CrossEntropyLoss	Weight update and error computation

IV. RESULTD AND DISCUSSION

A. *ResNet50 Model Evaluation Across Data-Split Scenarios*

The performance evaluation of the ResNet50 model was carried out across five training and validation data-split scenarios. For each scenario, the learning rate, batch size, and number of epochs that produced the best validation accuracy were selected. A summary of the evaluation results is presented in Table II and figure 3.

TABLE II

BEST RESNET50 RESULTS FOR EACH DATA-SPLIT SCENARIO

Data Split	Learning Rate	Batch Size	Epoch	Accuracy
50:50	0.0001	16	10	98.49%
60:40	**0.0001**	**16**	**30**	**99.22%**
70:30	0.0001	16	20	98.66%
80:20	0.0001	24	20	98.70%
90:10	0.0001	16	20	98.69%

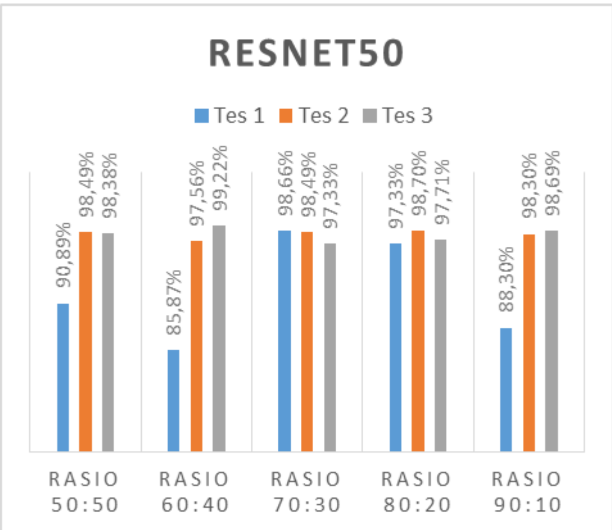

Figure 3 Accuracy comparison of the ResNet50 model under different train–test split ratios.

Based on Table II and Figure 3, the best performance of the ResNet50 model was achieved using the 60:40 train–test split with a learning rate of 0.0001, a batch size of 16, and 30 training epochs, resulting in an accuracy of 99.22%. This configuration also obtained a precision of 99.67%, a recall of 99.86%, and an F1-score of 99.26%. As illustrated in Figure 3, the 60:40 split consistently outperformed the other data-split scenarios across three experimental trials, demonstrating stable and reliable performance. These findings indicate that the ResNet50 model generalizes well to unseen data and effectively distinguishes between fatigued and non-fatigued facial images.

The ResNet50 architecture offers an advantage in feature extraction due to its deep network structure, which is capable of learning complex visual characteristics, such as eye conditions, facial expressions, and other visual cues associated with fatigue. The prediction results of the model are presented in Figure 4, demonstrating its ability to classify fatigued and non-fatigued conditions based on the learned visual features

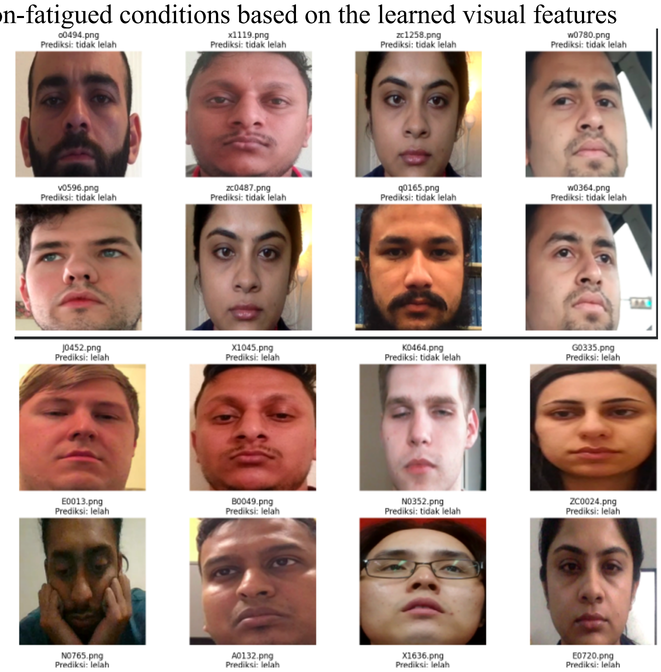

Figure 4. Prediction results of the proposed ResNet50 model.

B. *Training and Validation Curves*

Figure 4 and Figure 5 present the loss and accuracy curves of the best ResNet50 configuration (60:40 split, learning rate 0.0001, batch size 16, 30 epochs). The training and validation loss decrease sharply during the first ten epochs and continue to converge steadily until reaching a value close to 0.02 by epoch 30, with the validation curve closely tracking the training curve and only minor fluctuation, indicating that the model does not suffer from significant overfitting tabel.

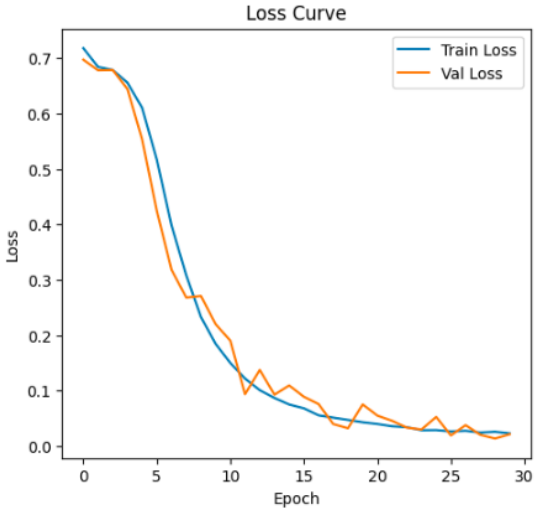

Figure 5. Training and validation loss curve of ResNet50.

Correspondingly, the accuracy curve in Figure 4 shows a rapid increase during the first ten epochs, after which both training and validation accuracy stabilize above 0.98 and continue to converge toward approximately 0.99 by the final epochs. The close alignment between the training and validation curves throughout training confirms that ResNet50 generalizes well to unseen validation data.

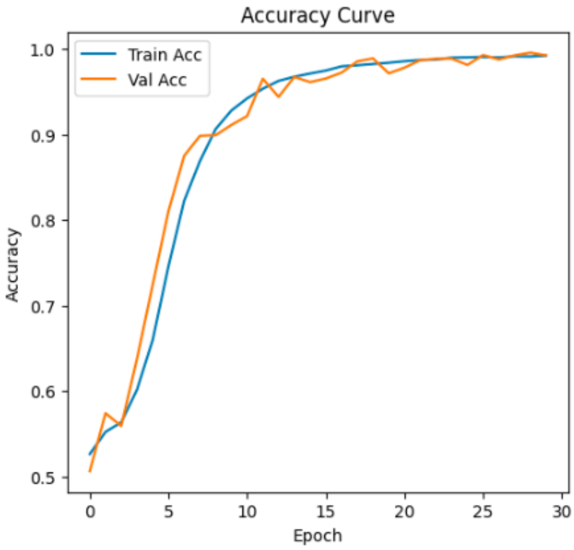

Figure 6. Training and validation accuracy curve of ResNet50.

C. *Confusion Matrix*

Figure 6 shows the confusion matrix produced by the best ResNet50 configuration on the validation data. Of the actual fatigued ("lelah") samples, 8,838 images were correctly classified as fatigued and 102 were misclassified as non-fatigued. Of the actual non-fatigued ("tidak lelah") samples, 7,749 images were correctly classified as non-fatigued and 29 were misclassified as fatigued.

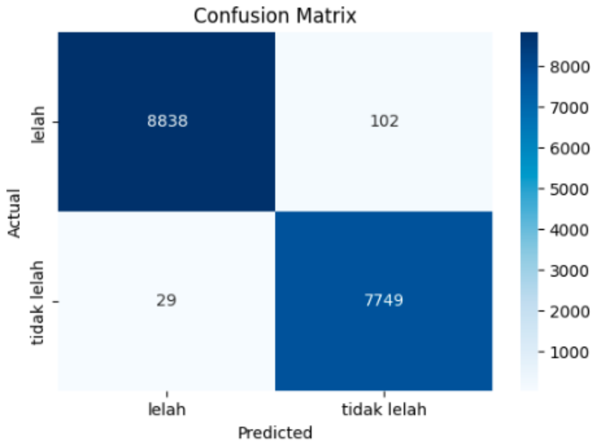

Figure 7. Confusion matrix of the ResNet50 model.

D. *Model Interpretability Using Grad-CAM*

Visual analysis using Grad-CAM was performed to understand the basis of the ResNet50 model's decisions when classifying images. Grad-CAM displays the image regions that contribute most to the prediction result in the form of a heatmap, as shown in Figure 7.

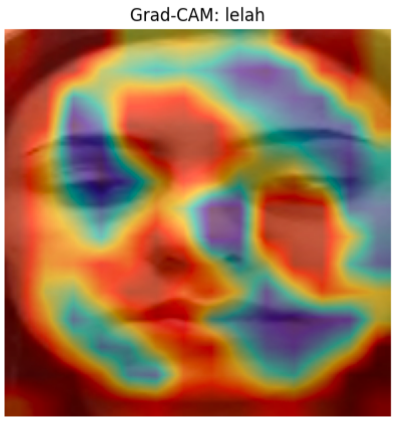

Figure 8. Grad-CAM Visualization of the ResNet50 Model

The visualization results show that the model focuses its attention on several important facial regions, namely the eyes, eyelids, under-eye area, and part of the cheeks and forehead. Regions with a high contribution to the model's decision are marked with warm colors such as red, yellow, and orange on the heatmap. The focus on the eye region indicates that the model is able to recognize visual characteristics related to fatigue, such as droopy-looking eyes or texture changes around the eyes. The residual connections in ResNet50 also make the feature-learning process more stable, allowing the model to better recognize complex visual patterns on the face rather than relying solely on the image as a whole.

E. *Grad-CAM Result Validation and Evaluation Using the Grad-CAM Result Dataset*

After obtaining the Grad-CAM visualizations, a validation step was carried out to determine whether the regions attended to by the model were consistent with the facial characteristics of fatigue. In addition to serving as an interpretation tool, the Grad-CAM visualization results were used to construct a new dataset consisting of Grad-CAM overlay images, which was subsequently used as input data to retrain the ResNet50 model. This retraining was performed to evaluate the effect of using the Grad-CAM-derived dataset on the model's performance.

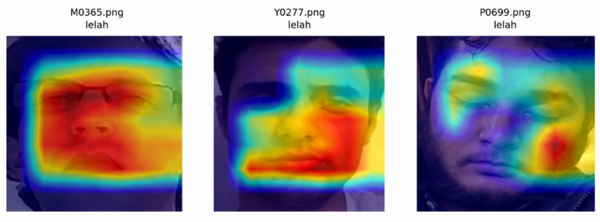

Figure 9. Prediction Results of the ResNet50 Model Trained on the Grad-CAM Result Dataset

Figure. 8 shows the prediction results of the ResNet50 model trained using the Grad-CAM-derived dataset on three test images that were successfully classified into the fatigued class. The Grad-CAM visualization shows that the model attends to the eye region as well as parts of the nose, cheeks, and mouth. For the first image, the model's attention remains spread across most of the face, whereas for the second image the focus is concentrated on the eyes, cheeks, and mouth. For the third image, the model's attention is more concentrated on both eyes, particularly around the eyelids. These results indicate that the model relies on a combination of several facial features, with the eye region as one of the most dominant contributors to fatigue detection.

To determine whether the regions attended to by the model correspond to facial fatigue indicators, the Grad-CAM visualization results were compared with eye detection using the Haar Cascade method. This comparison aims to assess whether the model's activation areas overlap with the eye region as one of the fatigue indicators, as shown in Figure 9.

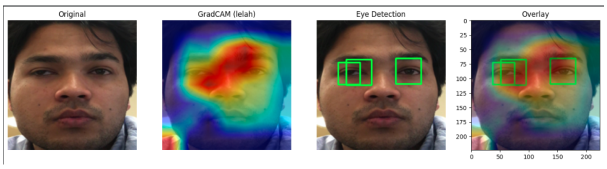

Figure 10 Comparison of Grad-CAM Visualization and Haar-Cascade Eye Detection

As shown in Figure. 9, the model successfully classified the image as fatigued. The Grad-CAM visualization shows that the model's activation area is located around the eyes, nose, and part of the forehead. The Haar Cascade eye-detection method successfully identified both eyes, allowing it to be used as a comparison against the model's activation area. After overlaying the two results, part of the Grad-CAM activation area overlaps with the eye region, although the activation also extends to the nose and forehead. This indicates that, for the visualized example, the model attends not only to the eye region but also to the nose and forehead, suggesting that the model's decision is influenced by a combination of several facial features.

F. *Model Implementation on a Real-Time Camera System*

The ResNet50 model with its best configuration was implemented in a camera-based system to detect fatigue conditions in real time. The camera serves as the input device that captures the user's facial image, which is then processed by the model to produce a rapid prediction in the form of a "fatigued" or "non-fatigued" label. Testing results show that the model not only performs well on the test data but can also be applied practically in a facial-image-based fatigue detection system, indicating its potential for use in monitoring applications with a high level of accuracy and efficiency.

V. CONCLUTION

This study successfully built and evaluated a Convolutional Neural Network (CNN) model based on the ResNet50 architecture to detect fatigue levels from facial expressions. The model was trained and validated across five data-split scenarios with varying learning rates, batch sizes, and numbers of epochs. The best performance of the ResNet50 model was obtained in the 60:40 data-split scenario with a harus diketik dengan font biasa *hypertext link* dan bagian *bookmark* akan dihapus. Jika paper perlu. learning rate of 0.0001, a batch size of 16, and 30 epochs, achieving an accuracy of 99.22%, a precision of 99.67%, a recall of 99.86%, and an F1-score of 99.26%.

The application of Grad-CAM successfully improved the interpretability of the ResNet50 model by highlighting the facial regions that most influence the classification result, particularly the eyes, cheeks, and forehead, thereby helping to explain the basis of the model's decisions and increasing trust in the predictions. In addition, the implementation of the model on a camera-based system showed that ResNet50 is able to provide fast and practical fatigue-condition predictions, indicating its potential for application in facial-image-based fatigue detection systems under real-world conditions.

Future research is recommended to use a dataset with a larger and more diverse range of samples, to develop the preprocessing stage further, such as through facial-landmark detection to focus the input on the eye region, and to develop the system into a real-time video-streaming form to make it more applicable in real-world use.

ACKNOWLEDGMENT

The authors would like to thank Ervin Yohannes, S.Kom., M.Kom., M.Sc., Ph.D., as the supervisor, for the guidance and direction provided throughout this study, as well as the Department of Informatics Engineering, Universitas Negeri Surabaya, for its research support and facilities.

REFERENSI

[1] R. Purwandari, A. Argiyanti, dan A. T. Afandi, “TINGKAT KELELAHAN DAN KECELAKAAN KERJA PERAWAT RUANG RAWAT INAP DI RUMAH SAKIT KABUPATEN JEMBER,” . *ISSN*, vol. 13, no. 2, 2024.

[2] A. S. Ritonga dan I. Muhandhis, "Analisis dan Implementasi Metode Viola-Jones dan CNN pada Sistem Deteksi Kantuk Real-Time," *JIC*, vol. 10, no. 2, hlm. 53–66, Des 2024, doi: 10.55635/jic.v10i2.227.

[3] D. Salem dan M. Waleed, "Drowsiness detection in real-time via convolutional neural networks and transfer learning," *J. Eng. Appl. Sci.*, vol. 71, no. 1, hlm. 122, Des 2024, doi: 10.1186/s44147-024-00457-z.

[4] B. Li dan D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, hlm. 57–64, Jun 2021, doi: 10.1016/j.ijcce.2021.02.002.

[5] I. C. Duta, L. Liu, F. Zhu, dan L. Shao, "Improved Residual Networks for Image and Video Recognition," 10 April 2020, *arXiv*: arXiv:2004.04989. doi: 10.48550/arXiv.2004.04989.

[6] H. Zhang dan K. Ogasawara, "Grad-CAM-Based Explainable Artificial Intelligence Related to Medical Text Processing," *Bioengineering*, vol. 10, no. 9, hlm. 1070, Sep 2023, doi: 10.3390/bioengineering10091070.

[7] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, dan M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, hlm. 92–110, Jul 2021, doi: 10.1016/j.neucom.2020.04.157.

[8] D. Nagaraju dan N. Chandrachoodan, "Compressing fully connected layers of deep neural networks using permuted features," *IET Computers & Digital Tech*, vol. 17, no. 3–4, hlm. 149–161, Jul 2023, doi: 10.1049/cdt2.12060.

[9] M. Krichen, "Convolutional Neural Networks: A Survey," *Computers*, vol. 12, no. 8, hlm. 151, Jul 2023, doi: 10.3390/computers12080151.

[10] S. H. Al-Gburi *dkk.*, "EffRes-DrowsyNet: A Novel Hybrid Deep Learning Model Combining EfficientNetB0 and ResNet50 for Driver Drowsiness Detection," *Sensors*, vol. 25, no. 12, hlm. 3711, Jun 2025, doi: 10.3390/s25123711.

[11] F. M. Ibrahim, R. N. Salmi, M. A. Saif, dan A. Mohammed, "Sleep Disorders' Prevalence and Impact on Academic Performance among Undergraduate Nursing Students in a Selected University, United Arab Emirates," *SAGE Open Nursing*, vol. 10, hlm. 23779608241274229, Jan 2024, doi: 10.1177/23779608241274229.

[12] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, dan B. Li, "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs," 19 Agustus 2020, *arXiv*: arXiv:2008.02312. doi: 10.48550/arXiv.2008.02312.

[13] Y. Albadawi, M. Takruri, dan M. Awad, "A Review of Recent Developments in Driver Drowsiness Detection Systems," *Sensors*, vol. 22, no. 5, hlm. 2069, Mar 2022, doi: 10.3390/s22052069.