Contents lists available at IdNNS
Journal of Intelligent Systems for Community Development
Volume 1 ; Issue 1
journal homepage: https://journal.idnns.org/index.php/jiscode

Image-Based Facial Skin Undertone Classification Using ResNet50 and MobileNetV2

Afina Putri Dzulqiyana[a,*], Herlawati Herlawati[a] and Andy Achmad Hendharsetiawan[a]

[a] *Informatics, Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya, Jl. Perjuangan No. 81, Marga Mulya, Bekasi Utara, Bekasi, 17143, West Java, Indonesia*

ARTICLE INFO

Keywords:
classification
skin tone
deep learning
ResNet50
MobileNetV2

ABSTRACT

This study aims to develop an automatic facial skin undertone classification system based on deep learning using facial images. The problem stems from the manual undertone identification process, which is subjective and potentially inconsistent when determining skin color categories. The method uses a Convolutional Neural Network (CNN) and compares two architectures, ResNet50 and MobileNetV2, for classifying warm, cool, and neutral skin undertones. The dataset combines secondary data from the Hugging Face platform and primary data obtained through direct facial image capture. The study follows the CRISP-DM process, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment. ResNet50 achieved an accuracy of 89.3% and a weighted F1-score of 0.893, whereas MobileNetV2 achieved an accuracy of 68.3% and a weighted F1-score of 0.684. Thus, ResNet50 demonstrated better and more stable performance for facial skin undertone classification.

1. Introduction

Facial skin color classification is an image-processing application that continues to develop alongside advances in Artificial Intelligence (AI). Classification aims to construct a model that recognizes and predicts data classes (Sihombing and Arsani, 2021). Information about skin undertones is important in cosmetics and fashion. However, undertone identification is still frequently performed manually, making the result subjective and potentially affected by lighting, camera quality, and image-capture angle (Prabowo et al., 2024). Therefore, a method capable of automatically recognizing image patterns is required, such as a Convolutional Neural Network (CNN). CNN is an extension of the Multi-Layer Perceptron (MLP) and is widely used in modern deep learning for image-classification problems (Santosa et al., 2025; LeCun et al., 1998).

Among various CNN architectures, two models widely used in applications are MobileNet and ResNet. ResNet (Residual Network) is a neural-network architecture introduced in 2015 by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ResNet uses residual blocks, enabling very deep networks to be trained without degradation in performance (Fitriasari and Rizkinia, 2021). With this structure, ResNet can improve accuracy as network depth increases, even to hundreds of layers. In contrast, MobileNet was designed to emphasize computational efficiency and low latency, making it suitable for mobile and embedded applications. MobileNet uses inverted residual blocks and depthwise separable convolution. These techniques work together to maintain model accuracy while reducing computational requirements, making MobileNet suitable for real-time mobile applications (Daviana and Anugraha, 2025).

In recent years, deep learning has been widely used for automatic image classification, but many studies have focused on general skin-color categories such as white, brown, and black (Nurapipah and Yuliana, 2025). Research specifically comparing ResNet50 and MobileNetV2 for facial skin undertone classification remains limited. Based on this problem, this study proposes an automatic image-based facial skin undertone classification system using CNN by comparing two deep learning architectures, ResNet50 and MobileNetV2. Classification is performed into three categories: warm, cool, and neutral. The objective is to examine the trade-off between efficiency and accuracy of each model and to recommend the CNN architecture most suitable for real-world image-based classification applications with either limited computational resources or high accuracy requirements.

2. Research Method

This study applies the Cross Industry Standard Process for Data Mining (CRISP-DM) as a guideline for structured

*Corresponding author: herlawati@ubharajaya.ac.id
ORCID(s):

data processing and model development. The method was selected because it covers the research process from problem identification through implementation. The focus is to compare the performance of ResNet50 and MobileNetV2 in classifying facial undertones into warm, cool, and neutral categories.

2.1. Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM is a universal data-mining methodology that is not dependent on a particular industry (Schröer et al., 2021). It is also widely used in data-mining and data-science projects because it provides a systematic and structured process (Mart et al., 2021).

2.2. ResNet50

ResNet50 is a deep-learning architecture based on the Convolutional Neural Network (CNN) introduced by Kaiming He through the work *Deep Residual Learning for Image Recognition* (He, n.d.). The model was designed to address the vanishing-gradient problem commonly encountered in very deep neural networks, enabling more stable training without degrading model performance. ResNet50 applies residual learning through shortcut or skip connections in each layer block. This mechanism allows the input to be passed directly to the output, so the model does not need to learn the complete transformation but instead focuses on the residual change. This approach can improve accuracy and accelerate convergence during training (Lin et al., n.d.).

The residual-learning concept in ResNet50 is expressed mathematically as

$$y = F(x, W_i) + x, \tag{1}$$

where x is the input from the preceding layer, $F(x, W_i)$ is the residual function learned by the network, and y is the final output after the input is added.

2.3. MobileNetV2

MobileNet is a CNN architecture developed by researchers at Google, including Andrew G. Howard and colleagues, and was introduced in the early development of the MobileNet family (Howard and Wang, 2012). The model remains CNN-based because it uses convolutional layers to extract image features. Its main distinction is the use of depthwise separable convolution, which separates convolution into two stages and significantly reduces the number of parameters and computational burden without a major loss of accuracy. Figure 3 illustrates the MobileNetV2 architecture, including the inverted residual blocks and the computational flow from input to output.

MobileNetV2 can be described mathematically by the following equation representing depthwise separable convolution:

$$y = DConv(x) + PConv(x), \tag{2}$$

where x is the input feature map, $DConv$ is depthwise convolution (convolution per channel), $PConv$ is pointwise convolution (1×1 convolution across channels), and y is the output feature map.

MobileNetV2 also uses an inverted residual connection to preserve feature information during network learning. The concept can be expressed as

$$y = x + F(x), \tag{3}$$

where x is the input feature, $F(x)$ is the transformation produced by the convolution block (inverted residual block), and y is the final output.

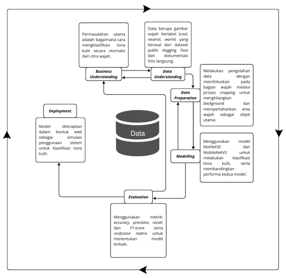


Figure 1: Research framework.

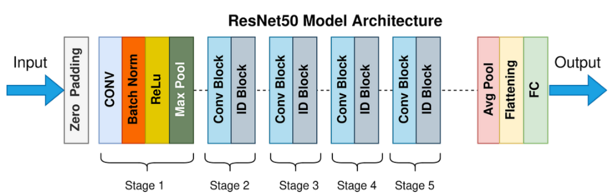


Figure 2: ResNet50 architecture (source: (Kanimozhi et al., 2024)).

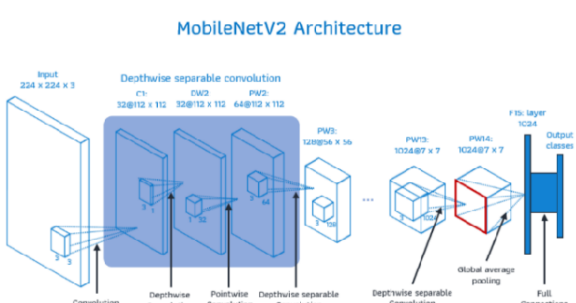


Figure 3: MobileNetV2 architecture (source: (Sandler et al., n.d.)).

3. Results and Discussion

This study uses a Convolutional Neural Network (CNN) approach with ResNet50 and MobileNetV2 to classify facial skin undertones into three categories: warm, cool, and neutral. The results and discussion are organized according to the CRISP-DM stages to describe the process systematically from data preparation to model evaluation.

Based on CRISP-DM, the process begins with problem understanding, collection of facial-image data from secondary sources, preprocessing, modeling, and evaluation using accuracy, precision, recall, and F1-score to compare the two architectures and determine the most optimal model. The selected model is subsequently implemented as a web application using the Flask framework for automatic skin-undertone classification.

3.1. Business Understanding

The objective is to develop a Deep Learning-based facial skin undertone classification system by comparing ResNet50 and MobileNetV2, with the expectation of obtaining accurate, consistent, and efficient classification. The results can support the fashion domain, particularly the selection of colors that match a user's skin undertone.

3.2. Data Understanding

The dataset combines two main data sources: secondary and primary data. Secondary data were obtained from the Hugging Face platform from the "undertone-dataset" owned by Nushiro. Primary data were obtained by collecting facial photographs from respondents through WhatsApp and storing them on Google Drive. The primary data contain human facial images with the same class specifications as the secondary data.

The secondary data consist of 1,094 images: 419 warm, 399 cool, and 276 neutral. The primary data consist of 30 images: 9 warm, 6 cool, and 15 neutral. The total dataset contains 1,124 facial images, comprising 429 warm images (38.08%), 409 cool images (36.03%), and 286 neutral images (25.89%). Table 1 presents the raw-data distribution.

Table 1
Raw data distribution.

Source	Class	Images	Percentage (%)
Secondary	Cool	399	35.50
Secondary	Neutral	276	24.56
Secondary	Warm	419	37.28
Primary	Cool	6	0.53
Primary	Neutral	15	1.33
Primary	Warm	9	0.80
Total	Cool	409	36.03
Total	Neutral	286	25.89
Total	Warm	429	38.08
Overall		1,124	100.00

3.3. Data Preparation

The data-preparation process consists of several stages.

1. Facial preprocessing. Facial preprocessing is performed before model training using a `FaceProcessor` class. The process includes illumination normalization using Color Constancy (Minkowski norm), face detection using BlazeFace Short Range (MediaPipe), and cropping of the central facial region (30–70%) focusing on important areas such as the cheeks, forehead, and nose. The images are then resized to 224×224 pixels. The overall preprocessing results are shown in Figure 5.

2. Data splitting after facial-area preprocessing. After preprocessing, the dataset is divided into training and validation sets using a stratified split with an 80:20 ratio. This method maintains the class distribution of cool, neutral, and warm in both subsets, making model training more stable and representative. The resulting distribution is shown in Table 2.

Table 2
Data distribution after cropping.

Class	Total	Training (80%)	Validation (20%)
Cool	405	324	81
Neutral	289	231	58
Warm	426	341	85
Total	1,120	896	224

3. Data augmentation. To increase training-data variation and reduce overfitting, augmentation is applied using the `ImageDataGenerator` from TensorFlow Keras. The techniques include horizontal flipping, rotation up to 15°, zoom of 0–20%, and brightness adjustment of 0.7–1.3. Augmentation is applied only to the training data, while validation data remain unchanged so that model evaluation remains objective.

3.4. Modeling

The modeling stage builds and trains two CNN architectures, MobileNetV2 and ResNet50, using ImageNet pretrained weights with all base-model layers frozen.

1. Model architecture. The model consists of an input layer, preprocessing layer, frozen base model as a feature extractor, Global Average Pooling 2D, a dense layer, a dropout layer, and a softmax output layer for three undertone classes: cool, neutral, and warm.

2. Hyperparameter settings. Hyperparameters were determined and tested incrementally to observe their effects on model accuracy and training performance. The configuration is presented in Table 3.

The training hyperparameters include learning rate, epochs, optimizer, batch size, and regularization techniques such as dropout, label smoothing, and L2 regularization.

Table 3
Hyperparameter configuration.

Hyperparameter	Setting
Learning rate	1×10^{-4} (0.0001)
Maximum epochs	50
Optimizer	Adam
Batch size	16
EarlyStopping patience	10
Label smoothing	0.1
L2 regularization	0.01
Dropout rate	0.4
Fine-tuning	No (frozen)

These parameters serve as the primary settings for optimizing facial skin undertone classification.

3. Model training. The hyperparameters are applied to the research dataset to improve model performance. The training results are presented in Figure 6.

As shown in Figure 6, MobileNetV2 stopped early at approximately epoch 20, with fluctuating validation accuracy between 65% and 75%, indicating less stable learning. In contrast, ResNet50 trained through the final epoch with more consistent results, reaching approximately 94% training accuracy and 89% validation accuracy, with lower and more stable loss than MobileNetV2.

3.5. Evaluation

After modeling, the test results are presented using a classification report and confusion matrix to evaluate the models for the cool, neutral, and warm classes. Table 4 summarizes the classification report.

ResNet50 achieved the best performance, with an accuracy of 89.3% and a weighted F1-score of 0.893. MobileNetV2 achieved a lower accuracy of 68.3% and a weighted F1-score of 0.684. Because ResNet50 produced the most optimal results, it was further analyzed using the confusion matrix in Figure 7 to examine the number of correct and incorrect predictions for each class.

Figure 7 shows a clear difference in performance between MobileNetV2 and ResNet50. For further clarification, the four evaluation metrics are defined below.

1. Accuracy. Accuracy is the proportion of correct predictions to the total number of samples. It is calculated as (Sathyanarayanan and Tantri, 2024)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

Here, TP denotes true positive, TN true negative, FP false positive, and FN false negative.

2. Precision. Precision measures the correctness of positive predictions and is calculated as (Sathyanarayanan and Tantri, 2024)

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{5}$$

3. Recall. Recall indicates the ability of the model to detect all positive samples and is calculated as (Sathyanarayanan and Tantri, 2024)

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{6}$$

4. F1-score. F1-score is the harmonic mean of precision and recall and balances the two metrics. It is calculated as (Sathyanarayanan and Tantri, 2024)

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

Figure 8 presents the comparison of accuracy and F1-score between ResNet50 and MobileNetV2 to facilitate identification of the best model for facial skin undertone classification.

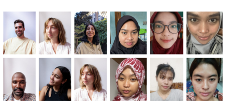
3.6. Deployment

In the deployment stage, the model is implemented as a practical application that provides recommendations such as clothing or hijab color selection. The study produces a web-based Undertone Classification Dashboard integrating ResNet50 and MobileNetV2 for real-time use. The system is built on CNN and uses the best trained model weights in `.h5` format. Figure 9 presents the dashboard interface and usage information.

Figure 10 shows the image-upload menu used for classification from an image uploaded by the user.

Table 4
Classification report.

Model	Class	Precision	Recall	F1-Score	Support
MobileNetV2	Cool	0.675	0.691	0.683	81
	Neutral	0.610	0.621	0.615	58
	Warm	0.744	0.718	0.731	85
	Macro Avg.	0.676	0.677	0.676	224
	Weighted Avg.	0.686	0.683	0.684	224
MobileNetV2 Accuracy		–	–	0.683	224
ResNet50	Cool	0.957	0.827	0.888	81
	Neutral	0.897	0.897	0.897	58
	Warm	0.844	0.953	0.895	85
	Macro Avg.	0.899	0.892	0.893	224
	Weighted Avg.	0.896	0.893	0.893	224
ResNet50 Accuracy		–	–	0.893	224


(a) Secondary dataset.


(b) Primary dataset.

Figure 4: Examples of the secondary and primary datasets.

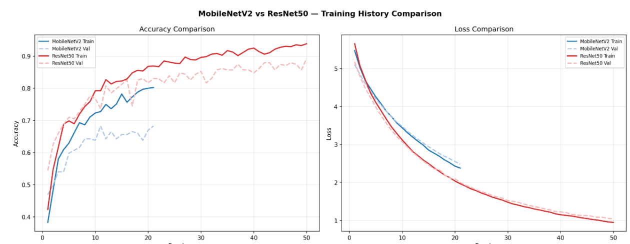


Figure 6: Training results.

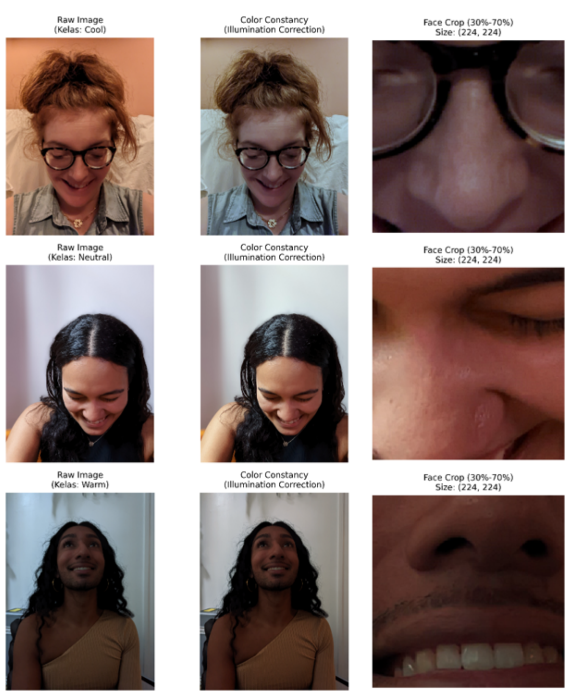


Figure 5: Facial preprocessing results.

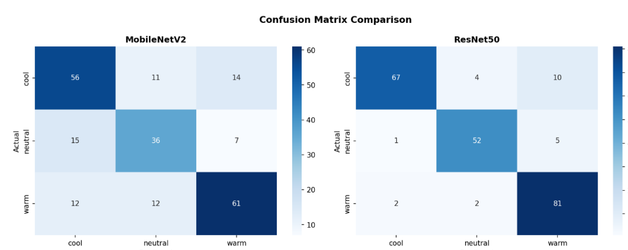


Figure 7: Confusion matrices for (a) MobileNetV2 and (b) ResNet50.

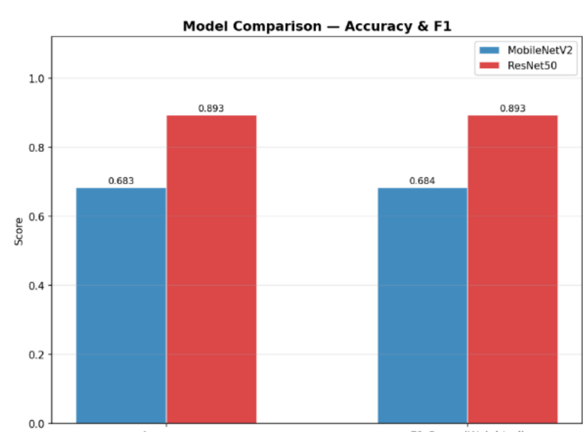


Figure 8: Comparison of accuracy and F1-score.

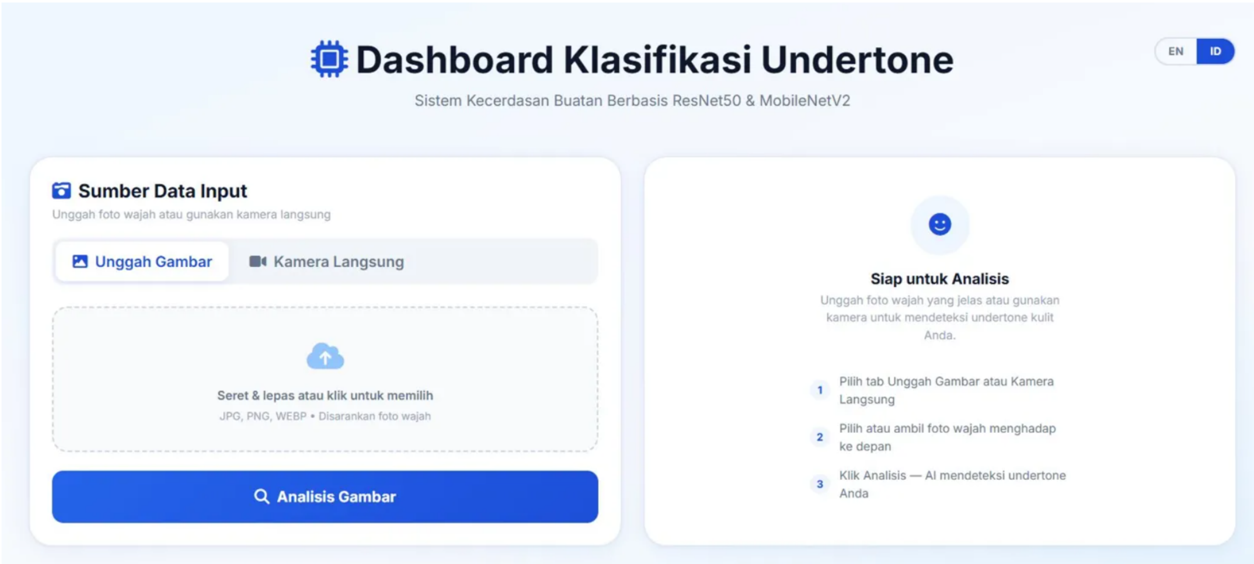
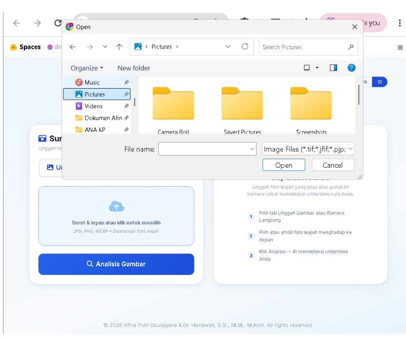
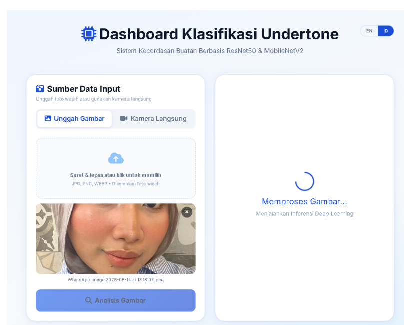


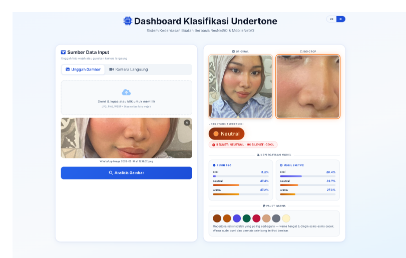
Figure 9: Undertone Classification Dashboard interface.



(a) Selecting an image.

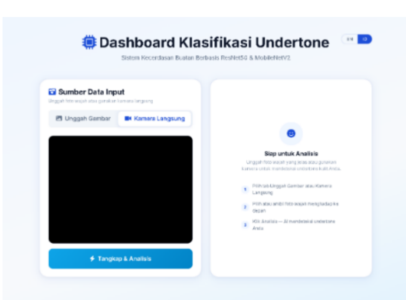


(b) Processing the image.

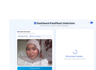


(c) Classification result.

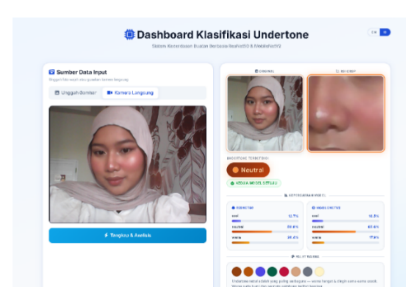
Figure 10: Image-upload classification interface: (a) image selection, (b) image processing, and (c) classification result.



(a) Initial camera menu.



(b) Image processing.



(c) Classification result.

Figure 11: Live-camera classification interface: (a) initial camera menu, (b) image processing, and (c) classification result.

4. Conclusion

This study successfully developed an image-based facial skin undertone classification system using a Convolutional Neural Network (CNN) and compared two architectures, ResNet50 and MobileNetV2. The evaluation showed that ResNet50 achieved the best performance, with 89.3% accuracy and a weighted F1-score of 0.893, while MobileNetV2 achieved 68.3% accuracy and a weighted F1-score of 0.684. These findings indicate that ResNet50 provides more stable and better feature extraction for facial skin undertone classification than MobileNetV2. In addition, the developed system was implemented as a real-time web application supporting both image upload and live-camera classification. Therefore, the resulting model provides not only an academic contribution but also potential practical value in fashion and beauty applications as a tool for recommending colors appropriate to a user's skin undertone. Future work can expand dataset diversity and optimize the model architectures to achieve better performance and greater adaptability to real-world data conditions.

Acknowledgments

The authors thank the reviewers for their comments and suggestions, which were helpful in improving the quality of this article. The authors also thank Universitas Bhayangkara Jakarta Raya for its support throughout the research process, including facilities, guidance, and opportunities to conduct and develop the research. This support played an important role in completing the study.

References

Daviana, F.P., Anugraha, N., 2025. Performance comparison between resnet50 and mobilenetv2 for indonesian sign language classification. JURIKOM (Jurnal Riset Komputer) 12, 319–328. doi:10.30865/jurikom.v12i3.8667.

Fitriasari, H.I., Rizkinia, M., 2021. Improvement of xception-resnet50v2 concatenation for covid-19 detection on chest x-ray images, in: 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), pp. 343–347. doi:10.1109/EIConCIT50028.2021.9431916.

He, K., n.d. Deep residual learning for image recognition. doi:10.1109/CVPR.2016.90.

Howard, A.G., Wang, W., 2012. Mobilenets: Efficient convolutional neural networks for mobile vision applications doi:10.48550/arXiv.1704.04861.

Kanimozhi, T., Rajeswari, V., Suguna, R., Nirmaladevi, J., Prema, P., Janani, B., Gomathi, R., 2024. Rwho: A hybrid of cnn architecture and optimization algorithm to predict basal cell carcinoma skin cancer in dermoscopic images. Scientific Temper 15, 2138–2142. doi:10.58414/SCIENTIFICTEMPER.2024.15.2.25.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. doi:10.1109/5.726791.

Lin, T., Ai, F., Doll, P., n.d. Focal loss for dense object detection. doi:10.1109/ICCV.2017.324.

Mart, F., Contreras-ochando, L., Lachiche, N., 2021. Crisp-dm twenty years later: From data mining processes to data science trajectories 33, 3048–3061. doi:10.1109/TKDE.2019.2962680.

Nurapipah, N., Yuliana, S.S., 2025. Skin tone classification in digital images using cnn for make-up and color recommendation. Journal of Intelligent Systems Technology and Informatics 1, 78–85. doi:10.64878/jistics.v1i2.29.

Prabowo, F.W., Homaidi, A., Lutfi, A., 2024. Deteksi warna kulit menggunakan metode deep learning dengan cnn (convolutional neural network). Jurnal Teknik Elektro Dan Informatika 19, 186–190.

Sandler, M., Zhu, M., Zhmoginov, A., Mar, C.V., n.d. Mobilenetv2: Inverted residuals and linear bottlenecks.

Santosa, R.R., Sasikirana, A., Ramadhan, Z.L., Putra, Y., Satria, B., 2025. Perbandingan kinerja mobilenetv2 dan resnet50v2 dalam klasifikasi tingkat kematangan tomat. Indonesian Journal of Science, Technology and Humanities 3, 49–56. URL: https://jurnal.intekom.id/index.php/ijstech.

Sathyanarayanan, S., Tantri, B.R., 2024. Confusion matrix-based performance evaluation metrics. AJBR 27. doi:10.53555/AJBR.v27i4S.4345.

Schröer, C., Kruse, F., Gomez, J.M., 2021. A systematic literature review on applying process models. Procedia Computer Science 181, 526–534. doi:10.1016/j.procs.2021.01.199.

Sihombing, P.R., Arsani, A.M., 2021. Perbandingan metode machine learning dalam klasifikasi kemiskinan di indonesia tahun 2018. Jurnal Teknik Informatika (JUTIF) 2, 51–56. doi:10.20884/1.jutif.2021.2.1.52.