Paradigma, Vol. 28, No. 1, March 2026, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 50-58
Published: LPPM Universitas Bina Sarana Informatika

PARADIGMA

website: http://jurnal.bsi.ac.id/index.php/paradigma/

Facial Skin Disease Classification Using Swin Transformer V2 and ResNet-50 in a Flask-Based System

Shinta Arum Imaniyah[1], Febrian Murti Dewanto[1], Nur Latifah Dwi Mutiara Sari[1]

[1]Informatics, University of PGRI Semarang, Semarang, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: 05-02-2026
Revised: 04-03-2026
Accepted: 16-03-2026
Available Online: 31-03-2026

Keyword:

Facial Skin Disease Classification
Medical Image Analysis
Swin Transformer V2
ResNet-50
Flask

ABSTRACT

Facial skin diseases are common health conditions that can significantly affect both physical and psychological well-being. Early identification is essential to minimize the risk of disease progression. However, in many areas, there is still a lack of access to dermatological care. Although deep learning algorithms have been widely used in medical image categorization, few studies offer a direct comparison between convolutional neural networks (CNN) and transformer-based architectures within a cohesive experimental framework, especially concerning the classification of facial skin diseases. This study compares the effectiveness of ResNet-50 with Swin Transformer V2 and develops a deep learning system to classify six different types of skin problems on the face. The models were evaluated using accuracy, precision, recall, and F1-score after the dataset was divided into subsets for testing, validation, and training. According to the trial results, Swin Transformer V2 achieves an astounding accuracy of 97.54%, outperforming ResNet-50, which achieves 94.44%. The training curves indicate stable learning behavior with minimal overfitting. Grad-CAM visualization is applied to improve interpretability by highlighting relevant regions in the images. The best-performing model is implemented in a Flask-based web application as a prototype system for early detection. These results demonstrate how transformer-based architectures can improve classification performance and highlight their potential applications in practical diagnostic support systems.

Corresponding Author:

Nur Latifah Dwi Mutiara Sari,
Informatics,
University of PGRI Semarang,
Jl. Sidodadi Timur No.24, Karang Tempel, Kec. Semarang Timur, Kota Semarang, Indonesia, 50232,
Email: nurlatifah@upgris.ac.id

INTRODUCTION

Skin diseases are among the most common health problems in Indonesia, with a relatively high prevalence ranging from 4.60% to 12.95% of the population (Gustiana, 2022). In urban areas such as Jakarta, non-infectious skin diseases such as dermatitis dominate patient visits, while acne (*acne vulgaris*) shows the highest prevalence among adolescent girls, reaching up to 85% (Agustin et al., 2024). Facial skin diseases not only cause physical discomfort but also have psychological impacts and may affect the quality of life of patients. Diagnosis is generally performed by dermatologists; however, the limited distribution of medical professionals and the relatively high cost of medical examinations make early detection still difficult to access (Wijaya et al., 2023). These drawbacks emphasize the need for more accessible, effective, and impartial technology-based diagnostic support solutions.

Medical image analysis, including the classification of skin diseases, has greatly improved thanks to recent developments in artificial intelligence, especially deep learning in computer vision (Jeong et al., 2022). Deep learning techniques show improved capacity to extract intricate visual information from medical photos as compared to traditional methods (Litjens et al., 2017). These improvements have enabled the development of automated systems that assist in early diagnosis and support clinical decision-making.

Convolutional Neural Networks are a popular deep learning technique for picture classification

DOI: https://doi.org/10.31294/p.v28i1.12381

applications because of its superior capacity to extract hierarchical information. Models such as ResNet-50 have demonstrated exceptional performance in a range of medical image classification applications, including dermatological image analysis (Zhong et al., 2024; Mahbod et al., 2021). These models are good at reproducing regional textures and patterns, which are crucial for recognising the outward signs of skin conditions.

To overcome this constraint, transformer-based designs such as the Vision Transformer were developed, incorporating self-attention mechanisms that provide global contextual modelling across visual patches (Khan et al., 2022; Dosovitskiy et al., 2021). Swin Transformer V2 presents an enhanced hierarchical architecture with shifting-window-based self-attention, facilitating more efficient computation while maintaining the capacity to capture both local and global features (Z. Liu et al., 2022). Prior research has shown that transformer-based models achieve competitive, consistent performance across many medical image processing tasks (Wei, Ren, Guo, Hu, & Liang, 2023), underscoring their significant potential for skin disease categorisation.

However, several research gaps remain. First, most existing studies predominantly focus on CNN-based models, with limited direct comparison between CNN and transformer-based architectures under a unified experimental framework, particularly for facial skin disease classification. Second, although transformer-based models have shown promising results, their comparative effectiveness in practical dermatological applications is still not sufficiently explored. Third, many previous studies emphasize model development and performance evaluation without integrating trained models into real-world systems that can be directly used by end users. These limitations indicate the need for a comprehensive study that not only compares different architectures but also evaluates their practical applicability (Mohan et al., 2024).

This study addresses these gaps by systematically comparing ResNet-50 and Swin Transformer V2 using an identical dataset and experimental settings. Furthermore, the best-performing model is implemented in a Flask-based web application to assess its usability as a diagnostic support tool. The research question of this study is: how do CNN-based and transformer-based architectures compare in terms of classification performance and practical applicability for facial skin disease detection ?. This study is expected to contribute both theoretically by providing a comprehensive comparison of deep learning architectures and practically by delivering an accessible prototype system for early detection of facial skin diseases.

RESEARCH METHOD

This study was conducted through several stages aimed at building, testing, and evaluating the performance of a facial skin disease classification model based on digital images.

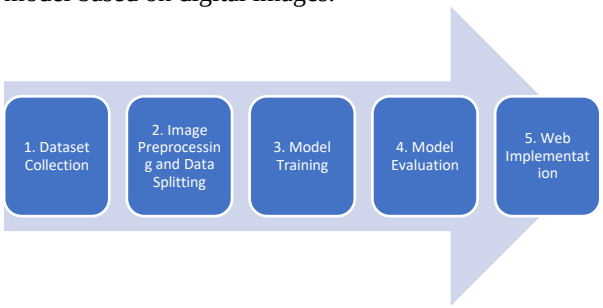

Figure 1. Research Stages

A Figure 1 shows that the study has several stages. First, data is collected and preprocessed. Next, the model is trained and tested. Finally, the model is deployed in a web-based application as a practical system.

1. Dataset Collection

This research utilizes the Augmented Skin Conditions Image Dataset developed by Syed Ali Raza Naqvi., which is available on the Kaggle platform. This dataset is balanced and consists of a total of 2,394 JPEG images (Naqvi, 2024).

The dataset is composed of six facial skin disease categories, including Acne, Carcinoma, Eczema, Keratosis, Milia, and Rosacea, with a total of 399 images for each category. The dataset is structured to be balanced across all classes to reduce potential bias during the model training process. The dataset is subsequently partitioned into training, validation, and test sets to facilitate objective evaluation of model performance.

2. Image Preprocessing and Data Splitting

The image preprocessing stage adjusts image quality and format to meet the model's requirements (Garcea, Serra, Lamberti, & Morra, 2023). To begin, all images are standardized to a resolution of 224 × 224 pixels before their values are normalized using the standard ImageNet mean and standard deviation (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Once these adjustments are made, data augmentation techniques such as rotation, horizontal flipping, scaling, and brightness adjustment are applied. These methods help make the data more diverse, improve the model's generalisation, and reduce the risk of overfitting during training (Mohan, Sivasubramanian, Sowmya, & Vinayakumar, 2024).

The application of data augmentation introduces additional variation in the dataset, enabling the model to capture more diverse visual features and improving its ability to avoid overfitting (X. Liu, Karagoz, & Meratnia, 2025).

Table 1. Image Dataset Distribution

Data Category	Number of Images
Training	1.675
Validation	359
Testing	360

The distribution of the dataset shown in Table 1 consists of 70% training data, 15% validation data, and 15% testing data. This arrangement separates the processes of model learning, hyperparameter adjustment, and performance evaluation, thereby supporting a fair and objective assessment.

3. Model Architecture Design

This research constructs a classification framework to investigate how different deep learning approaches perform in recognizing facial skin diseases from digital images. Within this framework, two architectures are implemented and compared to analyze their effectiveness in the classification task (Abdulqader & Abrahim, 2026). The two architectures used are ResNet-50 and Swin Transformer V2.

Swin Transformer V2 employs a hierarchical transformer architecture that integrates a shifting window self-attention mechanism to capture local features and broader contextual information in images (Z. Liu et al., 2022). Within this architecture, the input image is initially divided into non-overlapping patches, and thereafter processed via self-attention mechanisms within localised windows. The shifted-window approach facilitates interaction among neighbouring windows, permitting the model to acquire more extensive spatial linkages while preserving computational efficiency (Mohan et al., 2024). Transformer-based architectures have exhibited exceptional performance in numerous medical picture classification tasks due to their capacity to grasp long-range relationships and contextual patterns (Hatamizadeh et al., 2021).

Different from transformer-based approaches, ResNet-50 relies on a deep convolutional architecture with 50 layers, where residual connections are used to preserve gradient flow and prevent the vanishing gradient problem during training (Wei et al., 2023). Residual blocks allow the network to preserve information from earlier layers and ensure efficient gradient propagation during training. Owing to its effectiveness in hierarchical feature extraction, ResNet-50 has become a widely adopted architecture in medical image analysis, including dermatological image classification tasks. (Wei et al., 2023; Zhong et al., 2024).

The comparison between CNN and transformer architectures is necessary due to their distinct feature extraction strategies. CNNs specialize in learning local spatial patterns via convolution operations, while transformer models leverage self-attention to capture long-range contextual interactions across the entire image (Minaee et al., 2020). For this reason, the study evaluates and compares the performance of both architectures to identify the model that provides more accurate and robust results in facial skin disease classification.

Both models are designed and trained separately using the same dataset configuration and experimental settings to ensure a fair and objective comparison of their performance.

4. Training and Testing

The training process used the training dataset, where each image was first resized to a 224 × 224 input resolution (Wei et al., 2023). The models were trained for 20 epochs using a batch size of 16 with an initial learning rate of 1×10^{-4}. Parameter optimization was performed using the AdamW optimizer combined with the cross-entropy loss function (Loshchilov & Hutter, 2019).

The selection of hyperparameters was based on established configurations from prior research and initial tests to ensure consistent convergence and optimal performance. To improve the model's generalisation, various data augmentation techniques were employed during training, including random resizing, cropping, horizontal flipping, rotation, and colour jittering (Garcea et al., 2023). Additionally, the models were initialised using pretrained weights and executed via the PyTorch deep learning framework.

The training was conducted on a system using an NVIDIA RTX 2050 GPU, 16 GB of RAM, and an Intel Core i5 processor. GPU acceleration substantially improved training efficiency and reduced computational time.

Upon completing the training phase, the trained models were assessed on the test dataset to produce predictions for the six categories of facial skin diseases and to evaluate their overall classification performance.

5. Performance Evaluation

Model performance during training is monitored using the validation dataset to ensure that the learning process proceeds correctly. After the training stage is completed, the final evaluation is conducted on the testing dataset. The categorisation efficacy for each category of facial skin illness is subsequently assessed using measures including accuracy, precision, recall, and F1-score (Gupta & Kumar, 2021).

The classification performance is further analyzed using a confusion matrix, which illustrates how predictions are distributed between correct and misclassified classes. The formulas used to compute the evaluation metrics are shown in the following equations:

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

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

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

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

According to these equations, TP (True Positive) represents correctly predicted positive cases, TN (True Negative) denotes accurately predicted negative instances. FP (False Positive) refers to negative samples erroneously classified as positive,

while FN (False Negative) pertains to positive samples inaccurately classified as negative.

A confusion matrix is employed to illustrate classification outcomes for each class, displaying the quantity of accurate and erroneous guesses. To conduct a more in-depth analysis of the models' learning behaviour, the training accuracy and loss curves are also displayed. The graphs depict the convergence patterns of the models throughout training and help assess the stability of the learning process (Sokolova & Lapalme, 2021; Buda et al., 2021). In the concluding phase, the accuracy, precision, recall, and F1-score derived from both models are juxtaposed to evaluate their efficacy in categorising facial skin disease classifications.

6. Implementation

The trained and evaluated models are integrated into a web-based application using the Flask framework, enabling the deployment of deep learning models in a lightweight, flexible web environment (Sarker, 2021). Flask enables seamless integration with Python-based deep learning libraries such as PyTorch, enabling efficient machine learning inference within a web interface (Jayashree et al., 2025). Through this system, users can upload facial images and obtain automated predictions of facial skin diseases generated by the trained classification models (Orovwode et al., 2024).

Upon image upload, the system performs a series of preprocessing steps before model inference. The image is enlarged to the specified input resolution, its pixel values are normalised, and it is transformed into a tensor format compatible with the deep learning system. The processed image is subsequently assessed by the trained classification model, which generates probability outputs for each disease category; the category with the highest probability is designated as the final forecast (Rajaraman et al. 2018).

To make the system more transparent, it shows the predicted class and confidence score. It uses Grad-CAM (Gradient-weighted Class Activation Mapping) to highlight the parts of the image that most influence the model’s prediction (Selvaraju et al., 2019). The Grad-CAM heatmap is overlaid on the original image to show how the model makes decisions, highlighting the areas of the face that affect the classification result. This approach shows that the system can help as a computer-aided tool for early detection of facial skin diseases, but it still needs to be validated by professional dermatologists.

RESULTS AND DISCUSSION

This section presents the experimental results of the two facial skin disease classification models along with an in-depth analysis of the performance of the two architectures used. The evaluation was conducted to assess the models' ability to identify six classes of facial skin diseases based on digital images. In addition to presenting the experimental results, the model was also implemented in a Facial Skin Disease Classification System using a Flask-based framework, which is capable of predicting diseases accurately and rejecting inputs that are not facial images.

1. Model Testing Results

The experiments were conducted using two architectures, namely ResNet-50 and Swin Transformer V2, using 360 testing images consisting of six facial skin disease classes: acne, carcinoma, eczema, keratosis, milia, and rosacea. The effectiveness of the model was analyzed using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

Table 2. Performance Comparison of ResNet-50 and Swin Transformer V2 for Facial Skin Disease Classification

Model	Accuracy	Precision	Recall	F1-Score
Swin Transformer V2	0.9754	0.9757	0.9754	0.9752
ResNet-50	0.9444	0.9447	0.9444	0.9442

Table 2 shows that Swin Transformer V2 achieved 97.54% accuracy, surpassing ResNet-50 at 94.44%. The 3.1% enhancement signifies that transformer-based designs are superior in identifying intricate visual patterns in facial skin condition classification tasks. This finding aligns with prior research indicating that transformer-based models surpass CNN-based architectures in medical picture categorisation, owing to their capacity to capture global contextual information (Zhong et al., 2024).

The comparable precision and recall metrics for both models indicate that neither model demonstrates substantial bias towards any specific class. The superior F1-score of Swin Transformer V2 indicates a better balance between sensitivity and predictive accuracy, making it more reliable for real-world diagnostic applications.

The exceptional performance of Swin Transformer V2 is due to its self-attention mechanism, which allows the model to capture both local and global contextual information. In contrast to CNN-based architectures that focus on local feature extraction, transformer-based models can model long-range dependencies across multiple parts of an image. This capacity is especially crucial for classifying facial skin diseases, because visual characteristics may be distributed across several regions of the face.

These findings are consistent with previous studies. Mohan et al. (2024) reported that transformer-based architectures outperform CNN models in medical image classification tasks due to their ability to capture global contextual features. Similarly, Zhong et al. (2024) demonstrated that Swin Transformer provides improved performance in dermatological

image analysis compared to traditional CNN approaches.

2. Training Process Analysis

To evaluate the learning behavior of the models, an analysis of the training process was conducted by examining the changes in loss and accuracy on the training and validation datasets for Swin Transformer V2 and ResNet-50. These graphs were used to observe learning stability, performance consistency, and differences between training and validation performance.

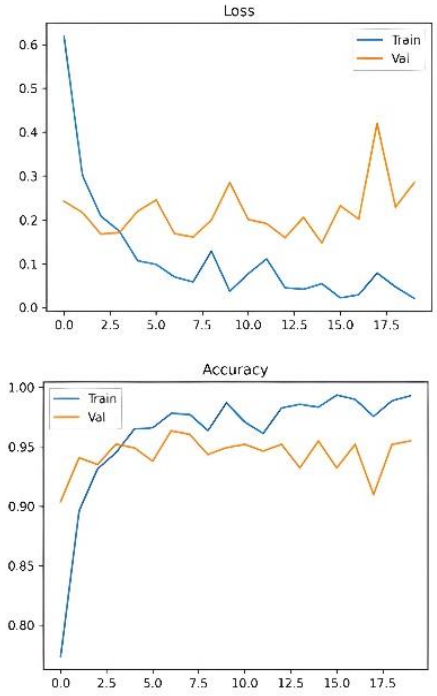

Figure 2. Training and Validation Loss and Accuracy Curves of Swin Transformer V2

According to Figure 2, the training loss declines continuously and markedly, reaching zero by the last period. This signifies that the model can effectively learn patterns from the training data (Rajpurkar et al., 2022). Nonetheless, the validation loss exhibits multiple oscillations during the intermediate to final phases of training and generally remains somewhat elevated compared to the training loss.

The discrepancies noted in the validation loss may be attributed to transformer-based models' susceptibility to data changes and their increased model complexity. Notwithstanding these variations, the rather narrow disparity between training and validation performance suggests that overfitting is minor and remains within an acceptable threshold.

Regarding accuracy, training accuracy rises rapidly to approximately 99%, whereas validation accuracy fluctuates between 95% and 97% with minor changes across epochs. The model attained a final accuracy of 0.9754, signifying exceptional classification performance. A minor discrepancy appears between training and validation accuracy; the gap is modest and remains within an acceptable range.

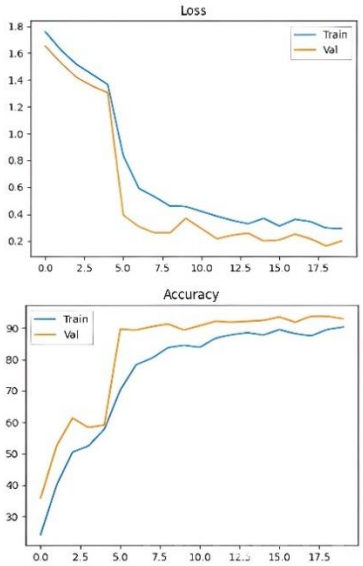

Figure 3. Training and Validation Loss and Accuracy Curves of ResNet-50

Based on Figure 3, both training loss and validation loss decrease consistently from the beginning of training until reaching relatively low and stable values in the final epochs. No significant spikes are observed in the validation curve, indicating that the training process proceeds in a stable manner.

The accuracy graph shows a gradual, steady increase in both training and validation accuracy. The ultimate accuracy of 0.9444 is obtained by closely matching the training accuracy with the validation accuracy. This trend suggests that the model's performance is constant across the training and validation datasets.

The more stable training behavior of ResNet-50 can be explained by its convolutional structure, which focuses on local feature extraction and typically requires fewer parameters compared to transformer-based models. Although it might restrict the model's capacity to incorporate global contextual interactions, this leads to more consistent convergence during training.

Overall, Swin Transformer V2 achieves higher accuracy than ResNet-50. However, ResNet-50 demonstrates a more stable training pattern, with a smaller gap between training and validation performance. This indicates that both models exhibit different learning characteristics, where transformer-based models provide higher representational capacity and accuracy, while CNN-based models offer more stable and computationally efficient training behavior. This trade-off between performance and stability is consistent with findings in previous studies on CNN and transformer architectures in medical image analysis.

These findings are consistent with previous studies showing that transformer-based models tend to

exhibit higher performance but are more sensitive to data variations compared to CNN-based architectures (Khan et al., 2022). This observation reinforces the trade-off between model performance and training stability, where transformer-based models offer higher accuracy while CNN-based models provide more stable and computationally efficient learning behavior.

3. Confusion Matrix Analysis

The distribution of accurate and inaccurate predictions for each class of face skin condition was assessed using confusion matrix analysis (Sathyanarayanan, 2024). This analysis provides deeper insight into class-wise performance and enables the identification of misclassification patterns across different categories.

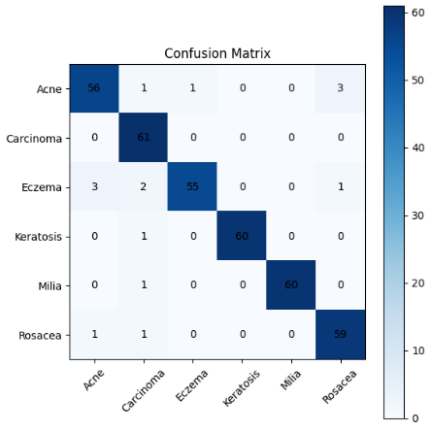

Figure 4. Confusion Matrix of Swin Transformer V2 for Multi-Class Facial Skin Disease Classification

Based on Figure 4, most values are concentrated along the main diagonal, indicating that the majority of the data are correctly classified. For the Acne class, the model correctly predicts 56 samples, with only a small number of errors in other classes. The Carcinoma class shows excellent performance with 61 correct predictions and almost no misclassification.

For the Eczema class, there are 58 correct predictions, while Keratosis and Milia each show 60 correct predictions. For the Rosacea class, the model produces 59 correct predictions, with only a minimal number of errors.

The consistently high number of correct predictions across all classes indicates that Swin Transformer V2 is highly effective in distinguishing between different facial skin disease categories. Its self-attention mechanism, which enables the model to capture both local and global contextual information, is responsible for this performance. As a result, the model is better able to differentiate subtle variations in visual patterns across facial regions, which are critical in medical image classification tasks.

Based on Figure 5, the main diagonal values also dominate, indicating that most samples are correctly predicted. For the Acne class, there are 53 correct predictions, with several misclassifications to other classes. The Carcinoma and Eczema classes each record 58 correct predictions, while the Keratosis class achieves 60 correct predictions.

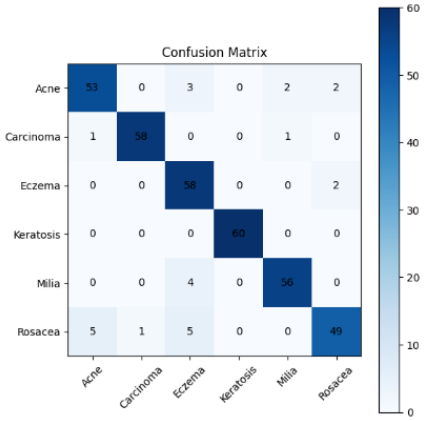

Figure 5. Confusion Matrix of ResNet-50 for Multi-Class Facial Skin Disease Classification

For the Milia class, there are 56 correct predictions with several misclassifications, and for the Rosacea class there are 49 correct predictions, showing a slightly higher number of errors compared to the previous model. The higher frequency of misclassifications in certain classes suggests that ResNet-50 faces challenges in distinguishing visually similar skin conditions. This restriction stems from CNN-based models' convolutional structure, which mostly concentrates on local feature extraction and might not adequately capture global contextual linkages throughout the image.

Overall, the comparison between the two models indicates that Swin Transformer V2 demonstrates superior classification performance with fewer misclassifications compared to ResNet-50. This finding highlights the advantage of transformer-based architectures in modeling complex visual dependencies. In contrast, although ResNet-50 provides relatively stable and reliable predictions, its performance is slightly limited when dealing with subtle inter-class variations.

This outcome is in line with other research showing that transformer-based models typically perform better than CNN models in medical picture classification tasks because of their capacity to capture contextual information and long-range relationships (Khan et al., 2022; Zhong et al., 2024). This result is also in line with previous studies such as Khan et al. (2022), which reported that transformer-based models consistently outperform CNN architectures in complex visual recognition tasks.

4. Web-Based System Implementation

The trained models were implemented in a web application developed with the Flask framework to assess their use as a diagnostic support tool for classifying facial skin diseases. This solution seeks to connect model development with practical application by enabling end users to interact directly with the trained system.

Users can upload facial pictures for examination via the web interface. After an image is

submitted, the system automatically carries out preparation tasks, including scaling the image to 224 × 224 pixels, normalising pixel values according to training parameters, and converting the image into a format compatible with the deep learning model. These preprocessing steps ensure consistency between training and inference data, which is crucial for maintaining model performance.

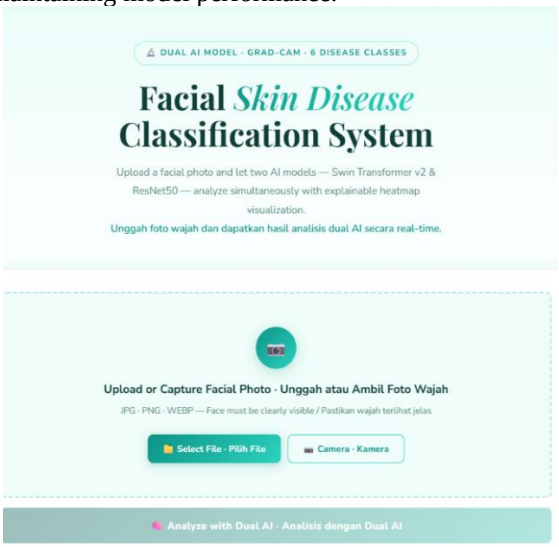

Figure 6. User Interface for Facial Image Upload in the Web-Based Classification System

Figure 6 illustrates that the system offers a straightforward and accessible interface for image uploads. Following preprocessing, inference is conducted using both ResNet-50 and Swin Transformer V2 models. These models generate probability scores for each class, with the final forecast chosen from the class having the highest likelihood. The probabilistic output enables the system to deliver both classification results and associated confidence levels, which are essential for decision support in medical applications.

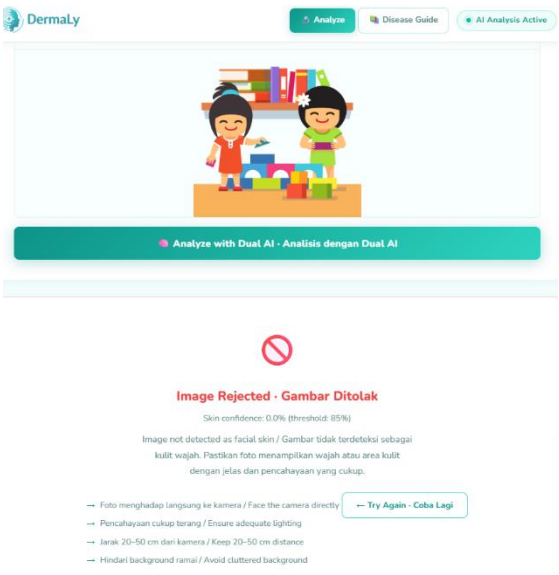

Figure 7. Invalid Image Detection and Validation Result Using MobileNetV2-Based Classifier

As shown in Figure 7, The system includes an input validation mechanism to ensure that only valid facial images are processed. This validation step is implemented using a MobileNetV2-based classifier to distinguish between facial and non-facial images. If the input image does not meet the required criteria, the system rejects the request and displays an error message. This mechanism improves system robustness and reduces the risk of incorrect predictions caused by invalid inputs.

Once the input image is validated, the classification process is performed, and the predicted disease category is displayed along with a confidence score. Grad-CAM (Gradient-weighted Class Activation Mapping) is used to create a heatmap visualisation that emphasises the areas most important to the model's prediction in order to improve interpretability. This feature is especially crucial for medical applications because it makes the model's decisions more transparent and easier for users to comprehend.

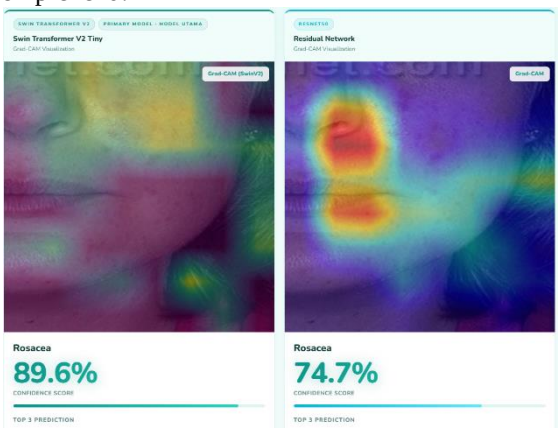

Figure 8. Classification Output with Predicted Label, Confidence Score, and Grad-CAM Visualization

The Grad-CAM heatmap visualisation and prediction results are shown in Figure 8, which also highlights the areas of the image that have the biggest impact on the model's choice. The inference process operates automatically and in real time with a short response time, which supports the system’s practical application.

The incorporation of deep learning models into a web-based platform indicates that the suggested solution is viable for real-world application and successful in experimental evaluation. The system serves as an accessible and interpretable tool for early detection of facial skin diseases. However, it is designed as a supportive tool and should not replace professional dermatological diagnosis. Future improvements may include integrating larger datasets and optimising inference efficiency to enhance scalability and reliability in clinical environments.

In conclusion, the incorporation of deep learning models into a web-based platform shows that the suggested approach is both feasible for real-world

deployment and beneficial in experimental evaluation. The system provides an accessible and interpretable tool for early detection of facial skin diseases. However, it is intended as a supportive tool and should not replace professional dermatological diagnosis. Future enhancements may involve incorporating larger datasets and optimising inference efficiency to improve scalability and reliability in clinical environments.

CONCLUSION

This study compares convolutional neural networks and transformer-based architectures for facial skin disease classification using digital image data and examines their deployment in a web-based diagnostic support system. The results indicate that transformer-based models possess enhanced capabilities for capturing complex visual patterns and global contextual information, making them particularly effective for medical image classification. The primary contribution of this work is the systematic evaluation of distinct deep learning paradigms within a unified experimental framework, along with the demonstration of their practical utility through an accessible, interpretable web-based application. The incorporation of Grad-CAM improves model transparency, which is critical for fostering trust in medical artificial intelligence systems. These findings underscore the potential of advanced deep learning methods to facilitate early detection of skin diseases and expand access to diagnostic tools, especially in regions with limited dermatological resources. Nevertheless, the study is constrained by the dataset's size and variability, which may limit generalizability to real-world scenarios. Future investigations should prioritise the inclusion of larger and more diverse datasets, enhancement of model robustness, and assessment of system performance in clinical settings to ensure reliability and scalability.

ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the Informatics Study Program, Faculty of Engineering and Informatics, Universitas PGRI Semarang, for the academic support and facilities provided during this research. The author also thanks the supervising lecturers for their valuable guidance and suggestions throughout the research process. In addition, the author appreciates the support and encouragement from family and friends during the completion of this study.

REFERENCES

Abdulqader, Z., & Abrahim, A. (2026). Deep learning-based skin disease detection and Classification: A Systematic Literature Review. *Dasinya Journal for Engineering and Informatics*, *2*(1). https://doi.org/10.65542/djei.v2i1.19

Agustin, D., Nurdini, R., & Noviyanti, L. (2024). Edukasi Pencegahan Dermatitis pada Lingkungan Pondok Pesantren Darul Huffaz Karawang. *Jurnal Kreativitas Pengabdian Kepada Masyarakat (PKM)*, *7*(8), 3446–3458. https://doi.org/10.33024/jkpm.v7i8.15424

Buda, M., Maki, A., & Mazurowski, M. A. (2021). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Garcea, F., Serra, A., Lamberti, F., & Morra, L. (2023). Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, *152*, 106391. https://doi.org/10.1016/j.compbiomed.2022.106391

Gupta, V., & Kumar, E. (2021). Review on Machine Learning Techniques for International Trade Trends Prediction. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 544–547. https://doi.org/10.1109/ICAC3N53548.2021.9725585

Gustiana, E. (2022). Hubungan Pengetahuan Tentang Personal Hygiene Dan Pemanfaatan Fasilitas Sanitasi Lingkungan Dengan Kejadian Penyakit Infeksi Kulit Pada Pondok Pesantren Anshor Al-Sunnah Air Tiris. *PREPOTIF Jurnal Kesehatan Masyarakat*, *6*(1), 1003–1007. https://doi.org/10.31004/prepotif.v6i1.26527

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1748–1758). https://doi.org/10.1109/WACV51458.2022.00181

Jayashree, B., Supriya, G., Chowdry, A. K., Professor, A., & Bahadur, R. Y. (2025). AI-Powered Patient Health Monitoring System Using Flask. International Journal of Advanced Research in Science, Communication and Technology International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal, 5(5). https://doi.org/10.48175/IJARSCT-27524

Jeong, H. K., et al. (2022). Deep learning in dermatology: A systematic review of current approaches and applications. Annals of Dermatology, 34(6), 434–445. https://doi.org/10.5021/ad.22.043

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3505244

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). *A Survey on Deep Learning in Medical Image Analysis*. https://doi.org/10.1016/j.media.2017.07.005

Liu, X., Karagoz, G., & Meratnia, N. (2025). Analyzing the Impact of Data Augmentation on the Explainability of Deep Learning-Based Medical Image Classification. *Machine Learning and Knowledge Extraction*, *7*(1). https://doi.org/10.3390/make7010001

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., … Guo, B. (2022). Swin Transformer V2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR52688.2022.01238

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1711.05101

Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., & Ellinger, I. (2021). Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics, 89, 101867. https://doi.org/10.1016/j.compmedimag.2021.101867

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Mohan, J., Sivasubramanian, A., Sowmya, V., & Vinayakumar, R. (2024). *Enhancing Skin Disease Classification Leveraging Transformer-based Deep Learning Architectures and Explainable AI*. https://doi.org/10.1016/j.compbiomed.2025.110007

Naqvi, S. A. R. (2024). *Augmented skin conditions image dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/syedalirazanaqvi/augmented-skin-conditions-image-dataset

Orovwode, H., Ibukun, O., & Abubakar, J. A. (2024). A machine learning-driven web application for sign language learning. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1297347

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., … Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, *2018*(4). https://doi.org/10.7717/peerj.4568

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Journal of Computer Vision (IJCV)*. https://doi.org/10.1007/s11263-019-01228-7

Sokolova, M., & Lapalme, G. (2021). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 58(1), 102407. https://doi.org/10.1016/j.ipm.2020.102407

Wei, C., Ren, S., Guo, K., Hu, H., & Liang, J. (2023). High-Resolution Swin Transformer for Automatic Medical Image Segmentation. *Sensors*, *23*(7). https://doi.org/10.3390/s23073420

Wijaya, D. A., Triayudi, A., & Gunawan, A. (2023). Penerapan Artificial Intelligence Untuk Klasifikasi Penyakit Kulit Dengan Metode Convolutional Neural Network Berbasis Web. *Journal of Computer System and Informatics (JoSYC)*, *4*(3), 685–692. https://doi.org/10.47065/josyc.v4i3.3519

Zhong, F., He, K., Ji, M., Chen, J., Gao, T., Li, S., … Li, C. (2024). Optimizing vitiligo diagnosis with ResNet and Swin transformer deep learning models: a study on performance and interpretability. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-59436-2