narra j

Original Article

Psoriasis severity assessment: Optimizing diagnostic models with deep learning

Aga Maulana[1,2], Teuku R. Noviandy[1], Rivansyah Suhendra[3], Nanda Earlia[4,5], Cita RS. Prakoeswa[6], Tara S. Kairupan[7], Ghifari M. Idroes[8], Muhammad Subianto[9] and Rinaldi Idroes[10]*

[1]Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia; [2]Department of Information Technology, Faculty of Science and Technology, Universitas Islam Negeri Ar-Raniry, Banda Aceh, Indonesia; [3]Department of Information Technology, Faculty of Engineering, Universitas Teuku Umar, Meulaboh, Indonesia; [4]Dermatology Division, Dr. Zainoel Abidin Hospital, Banda Aceh, Indonesia; [5]Department of Dermatology and Venereology, Faculty of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia; [6]Department of Dermatology and Venereology, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia; [7]Faculty of Medicine, Universitas Sam Ratulangi, Manado, Indonesia; [8]Department of Nuclear Engineering and Engineering Physics, Universitas Gadjah Mada, Yogyakarta, Indonesia; [9]Department of Statistic, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia; [10]Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia

*Corresponding author: rinaldi.idroes@usk.ac.id

Abstract

Psoriasis is a chronic skin condition with challenges in the accurate assessment of its severity due to subtle differences between severity levels. The aim of this study was to evaluate deep learning models for automated classification of psoriasis severity. A dataset containing 1,546 clinical images was subjected to pre-processing techniques, including cropping and applying noise reduction through median filtering. The dataset was categorized into four severity classes: none, mild, moderate, and severe, based on the Psoriasis Area and Severity Index (PASI). It was split into 1,082 images for training (70%) and 463 images for validation and testing (30%). Five modified deep convolutional neural networks (DCNN) were evaluated, including ResNet50, VGGNet19, MobileNetV3, MnasNet, and EfficientNetB0. The data were validated based on accuracy, precision, sensitivity, specificity, and F1-score, which were weighted to reflect class representation; Pairwise McNemar's test, Cochran's Q test, Cohen’s Kappa, and Post-hoc test were performed on the model performance, where overall accuracy and balanced accuracy were determined. Findings revealed that among the five deep learning models, ResNet50 emerged as the optimum model with an accuracy of 92.50% (95%CI: 91.2–93.8%). The precision, sensitivity, specificity, and F1-score of this model were found to be 93.10%, 92.50%, 97.37%, and 92.68%, respectively. In conclusion, ResNet50 has the potential to provide consistent and objective assessments of psoriasis severity, which could aid dermatologists in timely diagnoses and treatment planning. Further clinical validation and model refinement remain required.

Keywords: PASI, psoriasis, deep learning, skin disease classification, diagnostic models

Introduction

Psoriasis is a chronic autoimmune condition affecting millions worldwide, characterized by skin inflammation and scaling. The global prevalence of psoriasis is estimated to be around 2–3% of the population, with significant variations across different regions and ethnicities [1]. Epidemiological studies suggest that psoriasis affects individuals of all ages, with onset peaks in early adulthood and later midlife. Psoriasis is not merely a dermatological condition but a

Received: September 19, 2024 - Accepted: December 15, 2024 - Published online: December 19, 2024

systemic inflammatory disease that, if left untreated, can progress into more serious complications such as psoriatic arthritis, which affects approximately 30% of psoriasis patients [1]. Accurate assessment of psoriasis severity is crucial for effective treatment planning and monitoring disease progression. Traditionally, dermatologists have relied on visual inspection methods such as the Psoriasis Area and Severity Index (PASI) to evaluate the extent and severity of psoriasis [2]. However, these conventional approaches are inherently subjective and prone to human error, potentially leading to inconsistent diagnoses and suboptimal treatment strategies [3-5].

PASI score quantifies the severity of psoriasis and is used to monitor the disease progression [6]. In some countries, including Indonesia, the PASI score is used as a criterion to determine eligibility for insurance coverage of biologic therapy in severe psoriasis cases. Additionally, the score is integrated into the management protocols for patients presenting with severe symptoms [7]. The score is calculated based on four criteria, namely erythema, thickness, scaliness, and the proportion of body surface area involved. The severity is based on the weights corresponding to each body region, such as the head, upper extremities, lower extremities, and trunk [8,9]. However, PASI score is prone to unpredictability since it depends on the clinician's subjective assessment, potentially resulting in diagnostic discrepancies.

Several computer-aided diagnosis (CADx) systems have been developed to provide more objective assessments of psoriasis. A systematic review highlighted the application of machine learning techniques in evaluating and managing psoriasis through algorithms that analyze visual data [10]. Another study reported that the machine learning-based automated quantification of affected skin regions could minimize the variability in clinical evaluations [11]. A previous study has successfully introduced a convolutional neural network (CNN)-based system for the identification of psoriasis from clinical images [12]. A previous study showed that a CNN could achieve dermatologist-level accuracy in distinguishing between different types of skin lesions [17]. Similarly, another published study demonstrated the effectiveness of deep learning in grading the severity of atopic dermatitis, a condition with some similarities to psoriasis in terms of assessment challenges [5]. Collectively, these studies illustrate the growing role of CADx systems in overcoming the limitations of traditional PASI-based evaluations [10-12].

Artificial intelligence (AI) has shown increasing potential in psoriasis assessment, with advancements in segmentation, classification, and disease severity categorization [13,14]. Early studies in 2018 demonstrated the promise of AI, with a deep neural network (DNN) achieving 91% accuracy on 5,700 images and random forest outperforming other models on 676 images with an F1-score of 0.71 [18,19]. In 2020, a CNN applied to 187 images achieved 80% accuracy in predicting two psoriasis types [20]. More recent work has focused on advanced architectures and dataset augmentation. For instance, a 2021 study reported F1-scores of 0.926 for k-nearest neighbor (k-NN) and 0.986 for random forest on augmented datasets [8]. In 2022, a pre-trained VGG-19 model achieved 84.2% accuracy on a dataset of 172 normal skin and 301 psoriasis images [21]. The most recent advancements in 2023 included an EfficientNet-B0 model trained on 14,096 images, achieving 84.84% accuracy [22]. These studies highlight AI's growing role in diagnosing and classifying psoriasis by leveraging visual disparities in erythema, induration, desquamation, and body surface area involvement, which are also the basis for PASI scoring and disease severity categorization [15,16].

Despite significant advancements in AI applications for psoriasis assessment, several gaps remain. Comprehensive comparative studies evaluating multiple state-of-the-art deep learning architectures on the same psoriasis dataset are scarce, and most research focuses on binary classification or lesion segmentation rather than multi-class severity classification based on PASI scores. Additionally, the performance of models across varying severity levels and the challenges of overfitting and generalization remains underexplored. To address these gaps, this study evaluates five deep learning architectures—ResNet50, VGGNet19, MobileNetV3, MnasNet, and EfficientNetB0—selected for their unique design philosophies and potential for psoriasis severity classification. ResNet50's skip connections capture complex features; VGGNet19 excels at detecting fine-grained details; MobileNetV3 balances accuracy and efficiency; MnasNet uses neural architecture search for optimized performance; and EfficientNetB0 applies compound scaling for computational efficiency. These models are assessed for their ability to provide consistent and objective evaluations, aiming to advance AI-driven tools in dermatology and enhance clinical decision-making for psoriasis severity assessment.

Methods

Study design and dataset

A deep learning approach to classify psoriasis from images was employed, where the systematic process is presented in **Figure 1**. The first step involved collecting psoriasis images from publicly available datasets and medical sources, which were then processed by determining the region of interest (ROI) to focus on the affected areas of the skin. Subsequently, data labeling was performed to categorize the images based on different types or severity levels of psoriasis. The dataset underwent median filtering to reduce image noise and improve the quality of the input data for training. The images were then used to train several modified deep-learning models, including ResNet50, VGGNet19, MobileNetV3, MnasNet, and EfficientNetB0. During the tuning model phase, the hyperparameters were adjusted to optimize the performance of each model. After training, the models were tested on a separate set of images to evaluate their classification accuracy. Finally, the performance results were analyzed to determine which model achieved the best accuracy and overall performance in classifying the psoriasis images. The dataset was collected from a previously reported study, which consists of 27,153 images classified into ten skin disease categories [23].

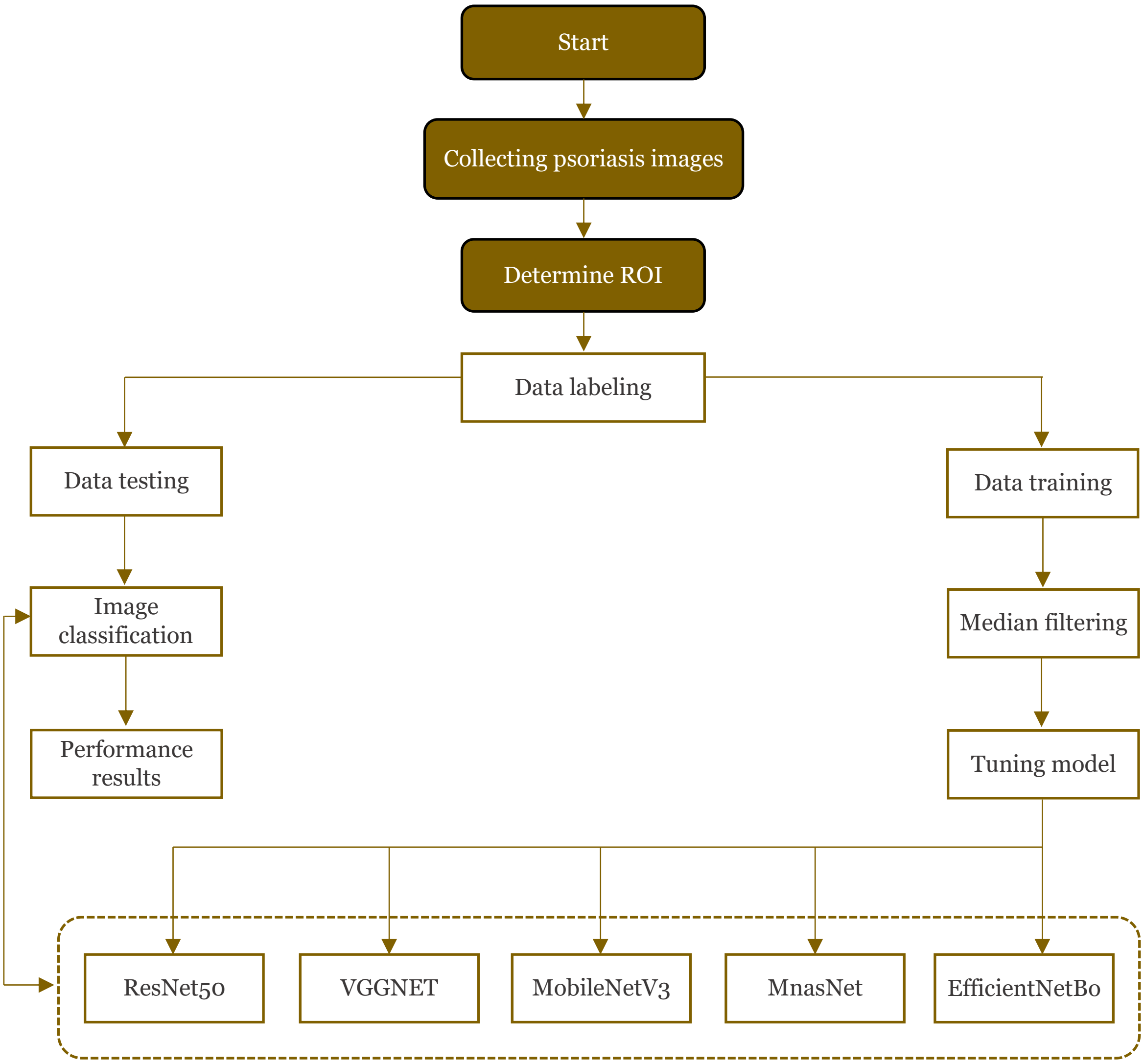

Figure 1. Research design flowchart for psoriasis severity classification models.

Image pre-processing

To focus on the affected skin areas, we employed several techniques. The ROI was determined using OpenCV, an open-source Python library, utilizing color and texture analysis techniques. Once the ROI was identified, the image was cropped to include only the affected area plus a 10% margin around it to ensure no relevant information was lost. All cropped images were then resized to a standard dimension of 250×250 pixels using bilinear interpolation to maintain consistency across the dataset. To reduce noise while preserving edges, a median filter was applied to the training data. The filter used a 3×3 pixels kernel size and a reflect 101 border type. The illustrations of the image modifications are presented in **Figure 2**.

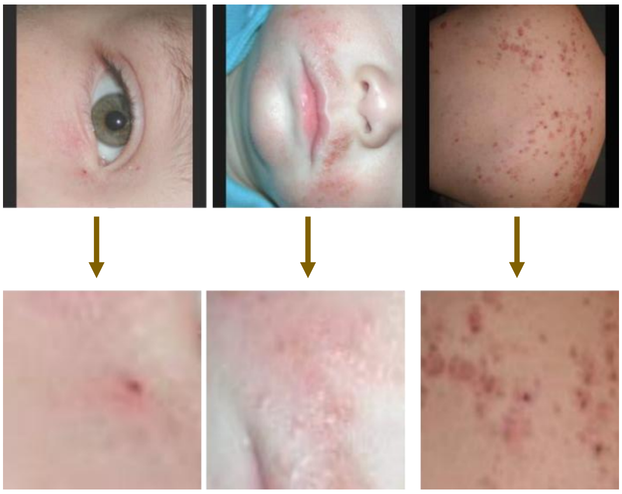

Figure 2. Representative of the cropping and resized technique used in this study.

Data labelling

The categorization of images into four severity classes (none, mild, moderate, and severe) was based on the PASI criteria. Each image was independently evaluated by three board-certified dermatologists, each with over ten years of professional experience, assigning scores for erythema, induration, and desquamation on a scale of 0–4 and estimated the affected area percentage for four body regions.

PASI scores were calculated using the standard formula, incorporating these assessments. Images were then classified as none (PASI=0), mild (PASI< 7), moderate (PASI 7–12), and Severe (PASI>12). During the labeling process, discrepancies between the dermatologists' classifications were resolved through a consensus-driven approach. Each conflicting case was thoroughly reviewed in a group discussion, with the dermatologists collectively examining the clinical features and PASI criteria to reach an agreement on the appropriate label. To ensure labeling consistency, 10% of the images were randomly selected for re-evaluation.

After these modifications, the final psoriasis dataset consisted of 1,546 images, categorized into four classes: "None" (490 images), "Mild" (247 images), "Moderate" (280 images), and "Severe" (529 images). The dataset was then split such that 70% (1,082 images) were allocated for training, while the remaining 30% (463 images) were used for validation and testing [24].

Proposed DCNN model

DCNN are a distinct category of artificial neural networks mostly used for image recognition, object identification, and classification applications. They play a significant role in deep learning, particularly for object identification and image analysis, and are widely applied in fields such as image processing, computer vision (including localization and segmentation), video analysis, autonomous vehicle obstacle detection, and speech recognition in natural language processing. DCNN have been widely adopted in various image classification tasks due to their adaptability and effectiveness. DCNNs have gained prominence in several picture categorization tasks due to their adaptability and efficacy. A standard DCNN architecture has many essential layers, each fulfilling a distinct function. This encompasses convolutional layers (such as conv2D and SeparableConv2D), pooling layers (such as MaxPooling2D and AveragePooling2), flatten layers, dropout layers, dense layers, and activation functions such as Rectified Linear Unit (ReLU). By intelligently integrating these layers, CNNs may be customized to do certain jobs with exceptional accuracy and efficacy.

In the present study, we developed a DCNN model by combining various types of layers. This model builds upon our previous work [5], with modifications that include the addition of a global average pooling 2D layer, a dense layer, a dropout layer, and the implementation of early stopping to enhance performance and prevent overfitting. To address the challenge of determining the optimal number of epochs, we used early stopping and trained the model for a maximum of 50 epochs. However, an early stopping mechanism was implemented, which halted the training process if the validation accuracy did not improve for five consecutive epochs.

Hyperparameters

In this study, the model was trained using specific hyperparameters that were carefully selected to optimize performance. An overview of the hyperparameters used during the training process is presented in **Table 1**. ReLU was used as the activation function for its efficiency in preventing the vanishing gradient problem and speeding up convergence. A batch size of 32 batches was chosen, balancing memory efficiency with model stability during training [25]. The learning rate was set to 0.001, which allowed the model to gradually adjust its weights without making overly large updates, thus preventing overshooting during optimization. The Adam optimizer was employed due to its adaptive learning rate capabilities [26]. Due to the involvement of multi-class classification, the categorical cross-entropy loss function was used in the model.

Table 1. Hyperparameters for training the models

Hyperparameter	Value
Activation	ReLU
Batch size	32
Learning rate	0.001
Optimizer	Adam
Lost function	Categorical cross-entropy

Evaluating the models

Diagnostic parameters such as accuracy, precision, sensitivity, specificity, and the F1-score were used to evaluate the model [27]. Given that psoriasis severity score is a multiclass classification issue, we calculated the weighted average of these measures to account for the relative representation of each class in the sample. This technique guarantees that each class's contribution is accurately represented in the final performance results. Mathematical expressions for determining accuracy, weighted average of precision, sensitivity, specificity, and F1-score are presented in **Equations 1–5**, following the suggestion from a previous study [5].

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \tag{1}$$

$$\text{Weighted average precision} = \frac{\Sigma\left(\frac{\text{TP}}{\text{TP} + \text{FP}}\text{i} * \text{weight}_{\text{i}}\right)}{\Sigma\,\text{weight}_{\text{i}}} \tag{2}$$

$$\text{Weighted average sensitivity} = \frac{\Sigma\left(\frac{\text{TP}}{\text{FN} + \text{TP}_{\text{i}}} * \text{weight}_{\text{i}}\right)}{\Sigma\,\text{weight}_{\text{i}}} \tag{3}$$

$$\text{Weighted average specificity} = \frac{\Sigma\left(\frac{\text{TN}}{\text{TN} + \text{FP}_{\text{i}}} * \text{weight}_{\text{i}}\right)}{\Sigma\,\text{weight}_{\text{i}}} \tag{4}$$

$$\text{Weighted average F1} - \text{score} = \frac{\Sigma(\text{F1} - \text{score}_{\text{i}} * \text{weight}_{\text{i}})}{\Sigma\,\text{weight}_{\text{i}}} \tag{5}$$

Where true positive (TP) is the count of positive cases that were correctly identified; false negative (FN) represents the number of cases erroneously classified as negative; false positive (FP) denotes the number of cases incorrectly labeled as positive; and true negative (TN) indicates the count of correctly identified negative cases [28].

Statistical analysis

To comprehensively evaluate the performance of the five deep learning models in classifying psoriasis severity, multiple statistical methods were employed. Pairwise McNemar's test was utilized to compare the classification performance of the models on the same dataset by examining the differences in misclassification rates. This test assessed whether the observed differences in performance between pairs of models were statistically significant, with *p*-values<0.05 considered indicative of significance.

Cochran's Q test was conducted to evaluate differences across all models simultaneously. This non-parametric test assessed whether the proportions of correct predictions differed significantly among the models. A significant Q statistic (*p*<0.05) indicated that at least one model's performance deviated from the others. Following this, post-hoc pairwise comparisons were performed using Dunn's test with Bonferroni correction to identify specific model pairs with significant differences, with adjusted *p*-values<0.05 considered significant.

To measure the level of agreement between model predictions and true labels, Cohen's kappa (κ) was calculated for each model. This metric quantifies classification consistency, with values ranging from 0 (no agreement) to 1 (perfect agreement). Higher κ values reflect better agreement, providing a robust measure of model reliability.

Heatmaps were generated to visualize the results of pairwise McNemar's tests. These visualizations offered an intuitive overview of the statistical significance of differences between model pairs, complementing the numerical analysis. Additionally, 95% confidence intervals for model accuracy were calculated using bootstrapping with 1,000 resamples. Non-overlapping confidence intervals between models further confirmed significant differences in performance.

Results

Models result

In this section, we successfully developed models for training psoriasis severity using a modified DCNN approach. All models have outstanding performance, indicating that our modified techniques successfully identified psoriasis severity. The confusion matrices across four psoriasis severity categories (none, mild, moderate, and severe) for MobileNetV3, ResNet50, VGGNet19, EfficientNetB0, and MnasNet are presented in **Figure 3**. The MobileNetV3 model correctly classified 111 and 97 cases belonging to Moderate and Severe categories, respectively. The model misclassified 12 Mild cases as None.

The ResNet50 model performed consistently well, particularly in classifying Moderate (n=109) and Severe (n=98) cases, with only minor misclassifications in the Mild category. Similarly, VGGNet19 demonstrated strong performance in classifying Moderate (n=111) and Severe cases (n=94) but showed limitations in distinguishing Mild from other severity levels, with several instances misclassified as None or Severe. EfficientNetB0 exhibited balanced performance across all categories, particularly for Moderate (n=107) and Severe cases (n=103), though it also encountered difficulties with the Mild category, similar to other models. In contrast, MnasNet exhibited the lowest accuracy, especially in the None and Mild categories, with notable misclassification of 17 Severe cases as Mild and 17 None cases as Severe.

Among the models (**Table 2**), ResNet50 achieved the highest performance with an accuracy of 92.50%, precision of 93.10%, sensitivity of 92.50%, specificity of 97.37%, and an F1-score of 92.68%. VGGNet19 followed with an accuracy of 90.00%, precision of 91.01%, sensitivity of 90.00%, specificity of 96.45%, and an F1-score of 90.27%. MobileNetV3 attained an accuracy of 89.38%, precision of 89.80%, sensitivity of 89.38%, specificity of 97.21%, and an F1-score of 89.46%. EfficientNet displayed similar performance, with an accuracy of 88.75%, precision of 89.21%, sensitivity of 88.75%, specificity of 96.89%, and an F1-score of 88.66%. MnasNet exhibited the lowest performance, with an accuracy of 66.87%, precision of 67.54%, sensitivity of 66.87%, specificity of 88.14%, and an F1-score of 66.89%.

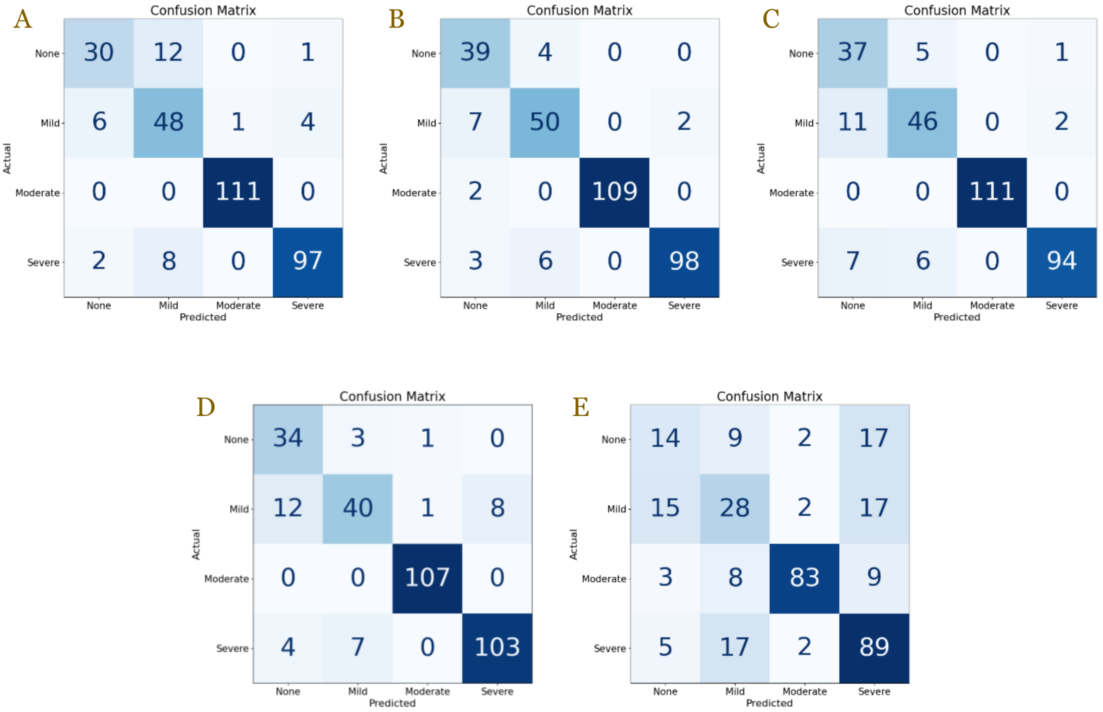

Figure 3. Confusion matrices for MobileNetV3 (A), ResNet50 (B), VGGNet19 (C), EfficientNetB0 (D), and MnasNet (E). The matrix compares the actual labels with the predicted labels. The rows represent the true classes, while the columns represent the predicted classes. The diagonal elements (top-left to bottom-right) indicate correct predictions, while off-diagonal elements show misclassifications.

Table 2. Performance result of modified MobileNetV3, ResNet50, VGGNet19, EfficientNet, and MnasNet models to determine the psoriasis severity

Model	Best epoch	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
MobileNetV3	17	89.38	89.80	89.38	97.21	89.46
ResNet50	17	92.50	93.10	92.50	97.37	92.68
VGGNet19	17	90.00	91.01	90.00	96.45	90.27
EfficientNet	10	88.75	89.21	88.75	96.89	88.66
MnasNet	6	66.87	67.54	66.87	88.14	66.89

The validation and training accuracy and loss curves for the models are presented in **Figure 4**. MobileNetV3 shows steady improvements in accuracy until the 10th epoch, followed by fluctuations in validation accuracy and an increase in validation loss. ResNet50 achieves steady and consistent improvements in both accuracy and loss, reaching 90% by the 10th epoch and continuing to improve without significant gaps. VGGNet19 shows stable validation accuracy around 90% but fluctuating validation loss between the 10th and 15th epochs. EfficientNetB0 improves in accuracy during the first 10 epochs, after which validation accuracy plateaus and loss fluctuates. MnasNet shows the weakest performance, with validation accuracy around 65–70% and stable loss, indicating underfitting.

Statistical analysis of model performance

To rigorously assess the performance differences among our five models, we conducted comprehensive statistical analyses. Pairwise McNemar's tests (**Table 3** and **Figure 5**) revealed statistically significant differences (p<0.05) between ResNet50 and all other models, confirming its superior performance. Specifically, the p-values for comparisons between ResNet50 and VGGNet19 (p=0.002), MobileNetV3 (p=0.001), EfficientNetB0 (p=0.001), and MnasNet (p<0.0001) were all below the significance threshold. This suggests that ResNet50 significantly outperformed these models.

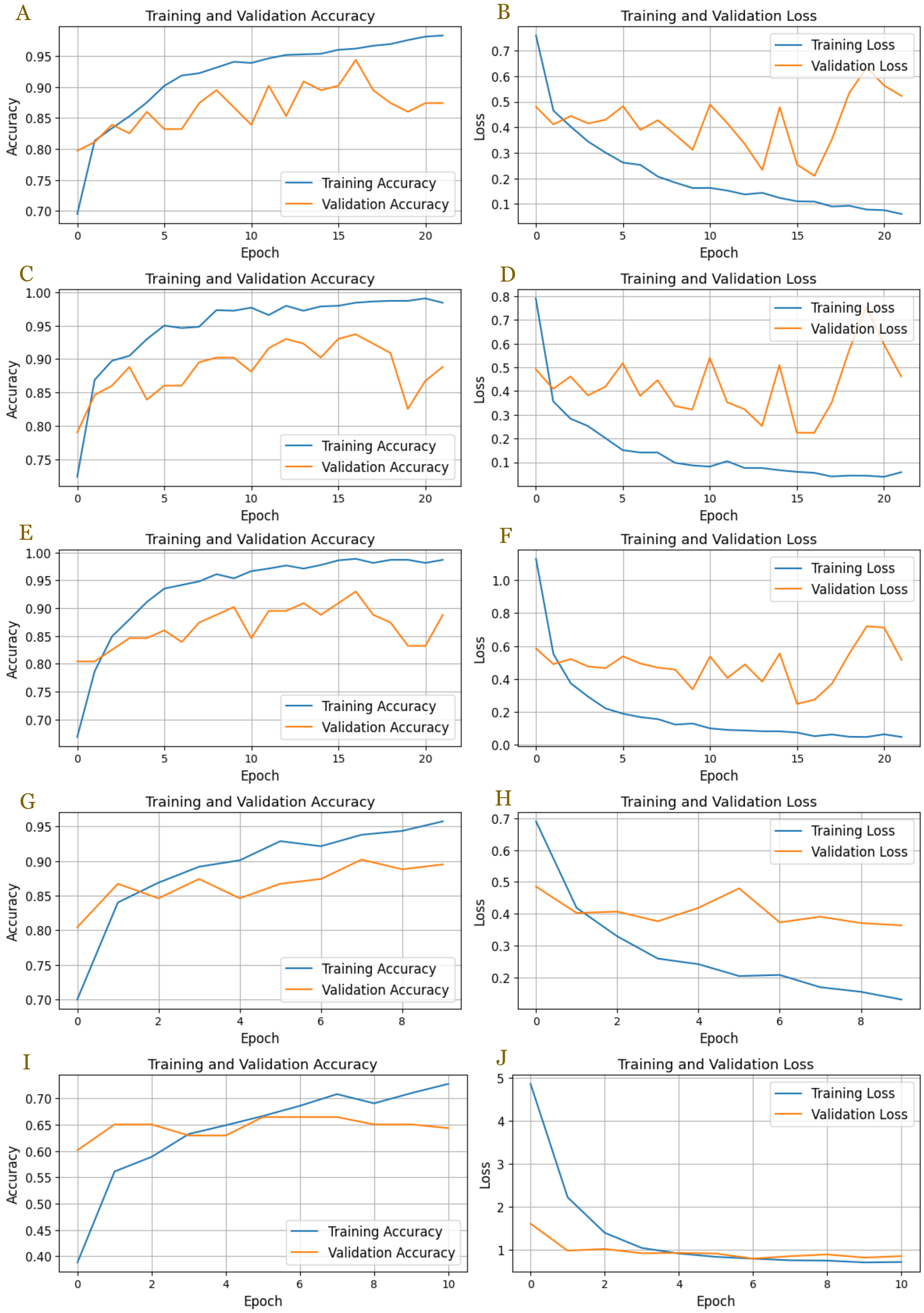

Figure 4. MobileNetV3 validation and training accuracy (A) and loss (B) curves. ResNet50 validation and training accuracy (C) and loss (D) curves. VGGNet19 validation and training accuracy (E) and loss (F) curves. EfficientNetB0 validation and training accuracy (G) and loss (H) curves. MnasNet validation and training accuracy (I) and loss (J) curves.

Additionally, MnasNet consistently underperformed across all pairwise comparisons, as indicated by its extremely low *p*-values ($p<0.001$). In contrast, there were no statistically significant differences between MobileNetV3 and EfficientNetB0 ($p=0.673$) or between VGGNet19 and MobileNetV3 ($p=0.187$), indicating similar performance between these models. A

Cochran's Q test showed significant differences across all models simultaneously (Q=78.3; df=4; p<0.001), indicating non-equivalent performances.

Post-hoc pairwise comparisons using Dunn's test with Bonferroni correction (**Table 4**) further confirmed ResNet50's significantly better performance (p<0.01 for all comparisons). This analysis provided a more detailed understanding of the relative performance of each model, reinforcing the standout capabilities of ResNet50 in our psoriasis severity classification task. We calculated 95% confidence intervals for each model's accuracy using bootstrapping with 1000 resamples. ResNet50's non-overlapping interval (91.2%–93.8%) further supported its superiority. This bootstrapping approach allowed us to estimate the reliability of our accuracy measurements and provided additional evidence for the robustness of ResNet50's performance. Cohen's kappa measurements showed ResNet50 achieved the highest agreement (κ=0.89) between predictions and true labels, while other models ranged from substantial (VGGNet19, κ=0.85) to moderate agreement (MnasNet, κ=0.58) (**Table 4**).

Table 3. Pairwise McNemar's test results (p-values)

Model	ResNet50	VGGNet19	MobileNetV3	EfficientNetB0	MnasNet
ResNet50	-	0.002	0.001	0.001	<0.001
VGGNet19	0.002	-	0.187	0.234	<0.001
MobileNetV3	0.001	0.187	-	0.673	<0.001
EfficientNetB0	0.001	0.234	0.673	-	<0.001
MnasNet	<0.001	<0.001	<0.001	<0.001	-

p-values<0.05 indicate statistically significant differences between models

Table 4. Model performance metrics and statistical analysis

Model	Accuracy (95% CI)	Cohen's Kappa (κ)	Post-hoc p-value*
ResNet50	92.50% (91.2–93.8%)	0.89	-
VGGNet19	90.00% (88.5–91.5%)	0.85	0.0089
MobileNetV3	89.38% (87.8–90.9%)	0.84	0.0042
EfficientNetB0	88.75% (87.1–90.4%)	0.83	0.0031
MnasNet	66.87% (64.8–68.9%)	0.58	<0.0001

Post-hoc p-values are from Dunn's test with Bonferroni correction, comparing each model to ResNet50

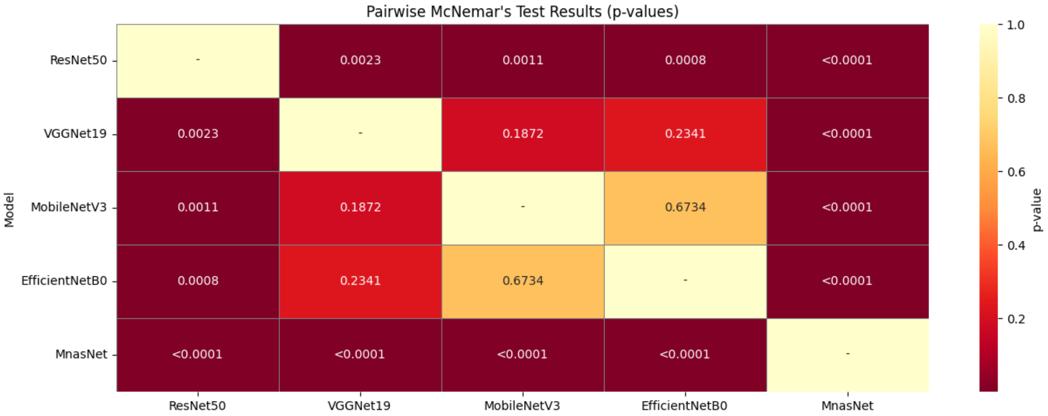

Figure 5. Heatmap for McNemar's test results.

Discussion

The present study revealed that ResNet50 is superior compared to other models, attributed to its deep architecture and skip connections, which allow for effective learning of complex features in psoriasis images. However, this performance advantage comes at the cost of increased computational complexity. The relatively poor performance of MnasNet, despite its neural architecture search optimization, suggests that the unique challenges of psoriasis severity classification may require more specialized architectures. Interestingly, all models have limitations in differentiating between mild and moderate cases, highlighting the subtle nature of these distinctions and the need for further refinement in feature extraction for these categories. The ResNet50 model used in the present study achieved higher accuracy (92.50%) than the Random Forest model (F1-score of 0.926) reported by Moon and colleagues in 2021 [8]. However, our results fall short of the 98.6% accuracy achieved by the EfficientNet-B0 model in another previous study [22]. This discrepancy could be due to differences in dataset size and composition, highlighting the critical role of diverse and comprehensive training data in determining model performance.

Based on the accuracy and diagnostic values achieved in this present study, the ResNet50 model demonstrates strong potential but might be insufficient to establish a clinical judgment in determining psoriasis severity. These values approach the thresholds typically emphasized in clinical guidelines, indicating reliability and consistency in high-stakes clinical settings. The evaluation of multiple models in the present study provides a comprehensive analysis of how different deep learning architectures handle psoriasis severity classification. The observed performance variations stem from the algorithmic differences among the models. For instance, ResNet50 utilizes residual connections, which enable the training of very deep networks by mitigating the vanishing gradient problem. This allows it to capture complex hierarchical features critical for differentiating subtle variations in psoriasis severity. In contrast, models like VGGNet19 rely on a sequential architecture with uniform layers, which, while effective for fine-grained features, may not match ResNet50's depth and ability to generalize across datasets. MobileNetV3 and EfficientNetB0 emphasize lightweight architecture and computational efficiency, balancing performance and resource utilization. MnasNet, developed through neural architecture search, optimizes for specific tasks but appears less effective in handling the nuances of psoriasis severity classification.

However, the findings from the present study also highlight several areas where future research should focus. One key observation is the tendency of some models to overfit during later epochs, indicating the need for more robust regularization techniques, such as early stopping or dropout layers, to improve generalization, as suggested in a previous study [32]. Additionally, while some models may underperform compared to others, their lower complexity may be beneficial in resource-constrained environments, highlighting the importance of balancing model complexity with accuracy depending on clinical use cases [33]. Future studies should explore optimizing these models with larger, more diverse datasets to ensure broader applicability across different skin types and patient populations, as limited datasets may restrict generalizability [34]. The integration of AI in dermatology, particularly in assessing psoriasis severity, represents a significant advancement in the field [29]. Recent studies demonstrate that deep learning models such as ResNet50, VGGNet19, and MobileNetV3 show considerable potential for accurately classifying psoriasis severity from image data, with ResNet50 often emerging as one of the most effective models [5,30]. These models can assist clinicians by providing consistent and objective assessments, reducing the variability that often arises from subjective visual evaluations [31]. The use of AI for psoriasis severity classification can improve diagnostic accuracy, streamline workflows, and ultimately enhance patient care by enabling faster, more personalized treatment decisions [15].

Several limitations of this study warrant attention. The dataset, while substantial, may not fully capture the diverse presentations of psoriasis across different skin types and ethnicities, potentially limiting model generalizability. Additionally, the use of 2D images excludes the textural details often critical in dermatological assessments, and reliance solely on visual features omits other clinical factors, such as patient-reported symptoms and quality of life impacts. Despite these limitations, this study highlights the potential of integrating advanced AI techniques, such as attention mechanisms to focus on relevant image regions and ensemble learning to combine model strengths, to enhance performance. Validation in real-world clinical settings and longitudinal studies to track disease progression are critical next steps. Furthermore, ensuring patient data privacy and addressing biases is essential for ethical and effective AI deployment. In conclusion, while this study demonstrates promising potential for AI in psoriasis severity assessment, further optimization, validation, and ethical implementation are essential for its clinical success.

Conclusion

Our study highlights the significant potential of AI, particularly ResNet50, in accurately assessing psoriasis severity with accuracy exceeding 90%. While this level of accuracy is promising, it may still be insufficient for high-stakes clinical applications in diseases like psoriasis, where misclassifications can impact treatment outcomes. Future research should focus on improving model robustness with larger, more diverse datasets and validating these tools in clinical settings. AI systems should be viewed as supportive tools to enhance, rather than replace, clinical expertise with ongoing efforts to address ethical considerations such as data privacy and bias.

Ethics approval

This study utilized publicly available online datasets of psoriasis images and did not involve human participants or identifiable private data. Therefore, ethical approval was not required for this research.

Acknowledgments

We would like to thank *Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi* and Universitas Syiah Kuala for supporting this study.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study was funded by *Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat - Direktorat Jendral Pendidikan Tinggi, Riset, dan Teknologi - Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi* and Universitas Syiah Kuala through "*Penelitian Fundamental Reguler*" Scheme with grant number: 658/UN11.2.1/PG.01.03/SPK/DRTPM/2024 for funding this research.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

How to cite

Maulana A, Noviandy TR, Suhendra R, *et al*. Psoriasis severity assessment: Optimizing diagnostic models with deep learning. Narra J 2024; 4 (3): e1512 - http://doi.org/10.52225/narra.v4i3.1512.

References

1. Michalek IM, Loring B, John SM. A systematic review of worldwide epidemiology of psoriasis. J Eur Acad Dermatol Venereol 2017;31(2):205-212.
2. Oakley A. PASI score. DermNet. 2009. Available from: https://dermnetnz.org/topics/pasi-score. Accessed: 17 May 2024.
3. Shrivastava VK, Londhe ND, Sonawane RS, *et al*. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: A first comparative study of its kind. Comput Methods Programs Biomed 2016;126:98-109.
4. Suhendra R, Suryadi S, Husdayanti N, *et al*. Evaluation of gradient boosted classifier in atopic dermatitis severity score classification. Heca J Appl Sci 2023;1(2):54-61.
5. Maulana A, Noviandy TR, Suhendra R, *et al*. Evaluation of atopic dermatitis severity using artificial intelligence. Narra J 2023;3(3):e511.
6. Fink C, Alt C, Uhlmann L, *et al*. Intra- and interobserver variability of image-based PASI assessments in 120 patients suffering from plaque-type psoriasis. J Eur Acad Dermatol Venereol 2018;32(8):1314-1319.
7. Evyana D, Novianto E, Budianti WK, *et al*. Association between the severity of hard-to-treat psoriasis and the prevalence of metabolic syndrome: A hospital-based cross-sectional study in Jakarta, Indonesia. PLoS One 2024;19(4):e0302391.
8. Moon CI, Lee J, Yoo H, *et al*. Optimization of psoriasis assessment system based on patch images. Sci Rep 2021;11(1):18130.
9. Fadzil MHA, Ihtatho D, Affandi AM, *et al*. Area assessment of psoriasis lesions for PASI scoring. J Med Eng Technol 2009;33(6):426-436.
10. Yu K, Syed MN, Bernardis E, Gelfand JM. Machine learning applications in the evaluation and management of psoriasis: A systematic review. J Psoriasis Psoriatic Arthritis 2020;5(4):147-159.

11. Meienberger N, Anzengruber F, Amruthalingam L, *et al*. Observer-independent assessment of psoriasis-affected area using machine learning. J Eur Acad Dermatol Venereol 2020;34(6):1362-1368.
12. Zhao S, Xie B, Li Y, *et al*. Smart identification of psoriasis by images using convolutional neural networks: A case study in China. J Eur Acad Dermatol Venereol 2020;34(3):518-524.
13. Ahmad A, Imran M, Ahsan H. Biomarkers as biomedical bioindicators: Approaches and techniques for the detection, analysis, and validation of novel biomarkers of diseases. Pharmaceutics 2023;15(6):1630.
14. Bhandari J, Awais M, Robbins BA, Gupta V. Leprosy. In: Ackley WB, Adolphe TS, Aeby TC, *et al*., editors. StatPearls. Treasure Island: StatPearls Publishing; 2024.
15. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-118.
16. Codella NCF, Gutman D, Celebi ME, *et al*. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington: IEEE; 2018.
17. Zhang N, Cai Y-X, Wang Y-Y, et al. Skin cancer diagnosis based on optimized convolutional neural network. Artif Intell Med. 2020;102:101756.
18. Syu J-M, Lai C-H, Lin G-S, Chai S-K. Psoriasis detection based on deep neural network. 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). Taichung: IEEE; 2018.
19. George Y, Aldeen M, Garnavi R. Psoriasis image representation using patch-based dictionary learning for erythema severity scoring. Comput Med Imaging Graph 2018;66:44-55.
20. Binti Roslan R, Mohd Razly IN, Sabri N, Ibrahim Z. Evaluation of psoriasis skin disease classification using convolutional neural network. IAES Int J Artif Intell 2020;9(2):349-355.
21. Aijaz SF, Khan SJ, Azim F, *et al*. Deep learning application for effective classification of different types of psoriasis. J Healthc Eng 2022;2022:1-12.
22. Huang K, Wu X, Li Y, *et al*. Artificial intelligence-based psoriasis severity assessment: Real-world study and application. J Med Internet Res 2023;25:e44932.
23. Hammad M, Pławiak P, ElAffendi M, *et al*. Enhanced deep learning approach for accurate eczema and psoriasis skin detection. Sensors 2023;23(16):7295.
24. Noviandy TR, Maulana A, Idroes GM, *et al*. QSAR-based stacked ensemble classifier for hepatitis C NS5B inhibitor prediction. 2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE). Banda Aceh: IEEE; 2023.
25. Noviandy TR, Maulana A, Zulfikar T, *et al*. Explainable artificial intelligence in medical imaging: A case study on enhancing lung cancer detection through CT images. Indones J Case Rep 2024;2(1):6-14.
26. Llugsi R, Yacoubi S El, Fontaine A, et al. Comparison between Adam, AdaMax and Adam W optimizers to implement a weather forecast based on neural networks for the Andean city of Quito. 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). Cuenca: IEEE; 2021.
27. Noviandy TR, Maulana A, Idroes GM, *et al*. An explainable multi-model stacked classifier approach for predicting hepatitis C drug candidates. Sci 2024;6(4):81.
28. Maulana A, Noviandy TR, Suhendra R, *et al*. Enhanced prediction of atopic dermatitis severity using advanced machine learning techniques. 2024 International Conference on Electrical Engineering and Informatics (ICELTICs). Banda Aceh: IEEE; 2024.
29. Tschandl P, Rinner C, Apalla Z, *et al*. Human–computer collaboration for skin cancer recognition. Nat Med 2020;26(8):1229-1234.
30. Han SS, Park GH, Lim W, *et al*. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2020;13(1):e0191493.
31. Brinker TJ, Hekler A, Enk AH, *et al*. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 2019;113:47-54.
32. Mahbod A, Schaefer G, Wang C, *et al*. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. Comput Methods Programs Biomed 2020;193:105475.
33. Liu Y, Jain A, Eng C, *et al*. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26(6):900-908.
34. Bissoto A, Fornaciali M, Valle E, Avila S. (De) constructing bias on skin lesion datasets. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE; 2019.