scientific reports

Check for updates

OPEN

Comparative performance of deep learning models and non-dermatologists in diagnosing psoriasis, dermatophytosis, and eczema

Nutcha Yodrabum[1], Chanisada Wongpraparut[2], Taravichet Titijaroonroj[3], Leena Chularojanamontri[2], Sumanas Bunyaratavej[2], Narumol Silpa-archa[2], Chayada Chaiyabutr[2], Thanapon Noraset[4], Teerapat Paringkarn[2], Thrit Hutachoke[2], Prameyuda Watchirakaeyoon[2], Pantaree Kobkurkul[2], Sirin Apichonbancha[1] & Praveena Chiowchanwisawakit[5]✉

Accurately differentiating scaly erythematous rashes among psoriasis, eczema, and dermatophytosis remains a clinical challenge, particularly for non-dermatologists. This study aimed to develop and evaluate deep learning models using macroscopic clinical images to classify these conditions and compare their performance with that of non-specialists. A total of 2940 images were sourced from public datasets, the Siriraj Dermatology databank, and newly collected images from Thai participants. Among sixteen evaluated models, the Swin demonstrated the best performance and interpretability. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations confirmed that the model focused on clinically relevant lesion features. Most importantly, in a pilot comparison, the Swin outperformed non-specialists in diagnostic accuracy. However, given the limited sample size of 30 images and 30 evaluators, these results should be interpreted as exploratory. Future studies with larger datasets and diverse clinician cohorts are warranted to confirm these findings and to support clinical integration.

Keywords Deep learning, Artificial intelligence, Psoriasis, Dermatophytosis, Eczema

Psoriasis, eczema, and dermatophytosis are common skin diseases characterized by erythematous papules or plaques with scales, frequently encountered in routine clinical practice[1]. Although these diseases have distinct etiologies, psoriasis is an immune-mediated disorder with keratinocyte hyperproliferation[2], eczema involving skin barrier dysfunction and immune dysregulation[3], and dermatophytosis caused by superficial fungal infection[4], they often present overlapping clinical features. This overlap contributes to diagnostic challenges, particularly among non-specialists, where misdiagnosis can result in inappropriate treatments that may exacerbate symptoms.[5].

Diagnostic accuracy in general practice remains limited, with some studies reporting accuracy rates as low as 50% for common skin diseases[6]. The increasing demand for dermatological care, especially in resource-constrained and rural settings, underscores the need for diagnostic tools that support non-specialists in clinical decision-making.

Recent advancements in artificial intelligence (AI) have demonstrated impressive capabilities in dermatologic image classification, often surpassing human performance in identifying skin cancers and other defined lesions[7–10]. For instance, AI systems have achieved up to 99% accuracy in differentiating melanoma from benign

[1]Division of Plastic Surgery, Department of Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [2]Department of Dermatology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [3]School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. [4]Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand. [5]Division of Rheumatology, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. ✉email: praveena.chi@mahidol.ac.th

lesions[7,11], and have performed well in detecting psoriasis and multiple skin diseases across various tasks[7,12]. However, accurate multiclass differentiation of erythematous scaly rashes, specifically psoriasis, eczema, and dermatophytosis, remains underexplored, with reported accuracies ranging from 89.1 to 96.2%[9,13,14].

Limitations of current models include their reliance on dermoscopic images and training on predominantly lighter skin phototypes, which may hinder generalizability to diverse populations and practical use in primary care[8]. In contrast, macroscopic clinical images captured via smartphones represent a more accessible format for real-world implementation, despite inherent quality variability. Leveraging AI to analyze such images could enhance dermatologic support in broader clinical environments.

This study addresses these gaps by developing and evaluating an AI framework to classify psoriasis, eczema, and dermatophytosis from macroscopic clinical images. We trained eight convolutional neural networks (CNNs) and eight Transformer-based models on a dataset of 2940 images from both public sources and Thai patients. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize model interpretability. Finally, we compared the diagnostic performance of our best-performing deep learning model with that of non-specialist clinicians to evaluate its practical utility.

Methods

The protocol for this study was approved by the Siriraj Institutional Review Board of the Siriraj Hospital of the Faculty of Medicine of Mahidol University (MU-MOU COA no. 073/2023). This study complied with the principles set forth in the Declaration of Helsinki of 1964 and all its subsequent amendments. The eligible skin lesion images were clinical images of scaly, erythematous rashes, with the final diagnosis of plaque psoriasis, eczema, or dermatophytosis (Figure 1). All patients contributing new images provided written informed consent for the use of their images in research, academic publication, and anonymized data sharing. Skin lesions on the face, neck, and groin, as well as tattoos or scars, were excluded due to the Personal Data Protection Act, which has been enforced in Thailand since June 2022.

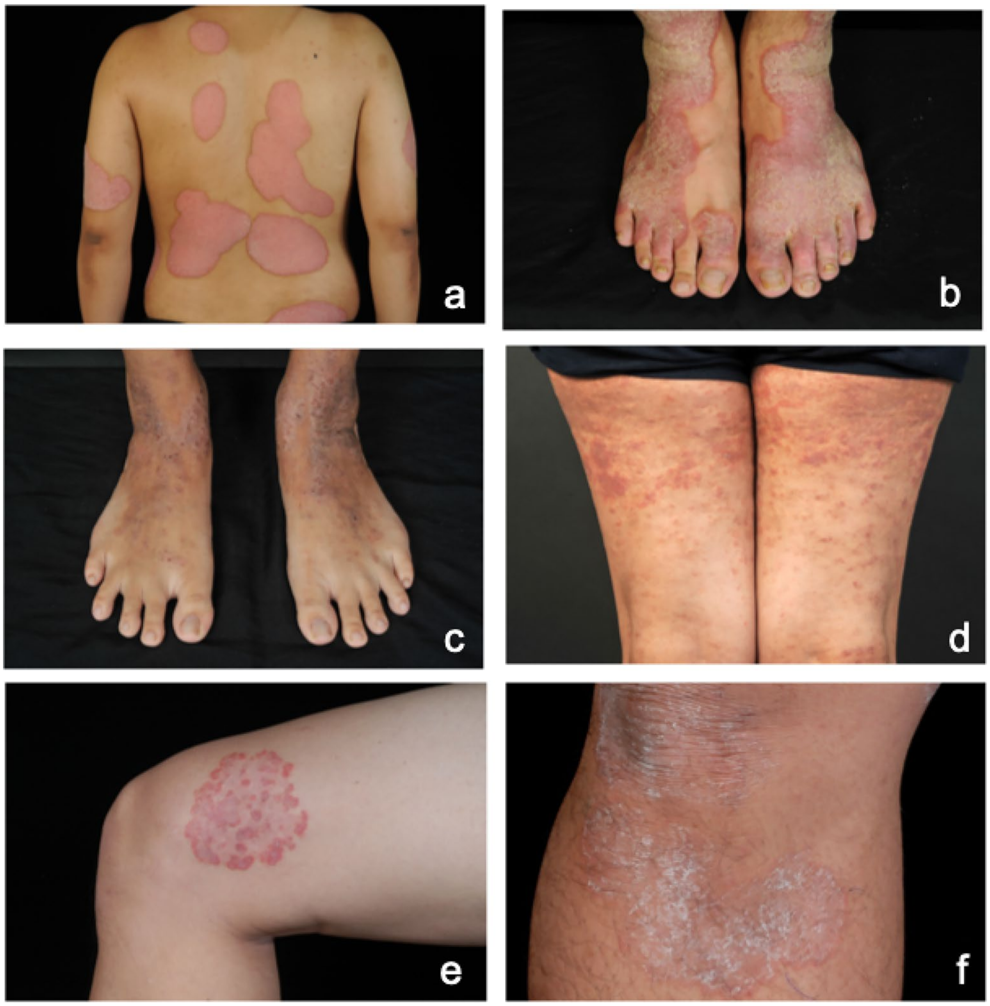

Fig. 1. The samples included three disease classes: (**a**, **b**) psoriasis, (**c**, **d**) eczema, (**e**, **f**) dermatophytosis, from Thai participants, Siriraj Dermatology Databank, Department of Dermatology, Faculty of Medicine Siriraj Hospital, Mahidol University.

Data acquisition

A total of 2940 photographs were included in this study, sourced from two primary categories: existing databanks and newly collected patient images. The existing databanks contributed 1320 images, comprising 308 images from the Siriraj Dermatology databank and 1012 images from public repositories, including DermNet[15]. The Siriraj Dermatology databank comprises images of Thai participants diagnosed with a range of skin diseases by experienced dermatologists employing state-of-the-art diagnostic methods. Participants provided informed consent for their images to be used in educational, research, and publication contexts. Dermatophytosis cases in this databank were confirmed through clinical manifestations and positive potassium hydroxide (KOH) examinations showing branching and septate hyphae.The diagnosis of psoriasis and eczema in the Siriraj Dermatology databank were also made by three experienced dermatologists, each with over 10 years of expertise in the field. For the public databank, the images were also reviewed and confirmed by the same three dermatologists. Notably, the combined dataset captures a broad spectrum of image characteristics—such as variation in quality, perspective, skin tone, and acquisition protocols. In the context of computer vision, training deep learning models on such diverse conditions can enhance model robustness, as it encourages generalization across real-world scenarios. Accordingly, the inclusion of diverse image sources may contribute to the stability and reliability of the model's performance across a wide range of clinical settings and input variations.

The newly collected dataset included 1620 images, obtained from Thai participants aged 18 years or older with scaly erythematous rashes and a confirmed diagnosis of psoriasis. Participants were recruited from the outpatient clinic at Siriraj Hospital and provided informed consent before inclusion in the study. Lesions were photographed using three different smartphone models: iPhone 11, 13, and 14 Pro (Apple Inc., Cupertino, CA, USA); Samsung Galaxy A33 (Samsung Electronics Co., Ltd., Suwon, South Korea); and Oppo A78 (Guangdong Oppo Mobile Telecommunications Corp., Ltd., Dongguan, China). Images were taken under consistent ambient lighting with a neutral green background and fixed distance (30 cm) to ensure reproducibility. Device flash was used under controlled conditions to enhance lesion detail without overexposure. Autofocus and exposure-lock features were used to maintain image sharpness and consistency across participants (Figure 2).

For each lesion, 18 photographs were captured: three angles (frontal, 30° left, 30° right) under both flash and non-flash conditions, with duplicate shots for quality assurance. From this set, one representative image per lesion was selected for inclusion, based on clarity, color balance, and lesion visibility. This selection was conducted by three dermatologists with over 10 years of clinical experience, using a consensus process to ensure diagnostic quality and consistency. An overview of the dataset is presented in Table 1.

Data preparation and data augmentation

To ensure that the training data were standardized, diverse, and suitable for effective training of both CNN and Transformer models, four steps of data preparation and augmentation (as shown in Fig. 2) were applied as follows. To minimize inter-device color variability and ensure consistency, all images underwent pixel intensity normalization to zero mean and unit variance, a standard procedure in deep learning workflows. This process adjusts brightness and contrast automatically without altering clinical features. No manual color correction or enhancement was applied. Although Figure 2 illustrates natural color variation due to lighting and skin tone, normalization ensured that models were not biased by these differences.

Step 1: Zero-padding to square image

To ensure consistency in input dimensions, we applied zero-padding to convert all images to a square shape. Zero-padding involves adding rows or columns of zeros around the image to make it square without altering the original content. This step helps maintain the aspect ratio and prevents distortion when resizing images later[16].

Step 2: Random horizontal flip

Random horizontal flipping with a probability of 0.5 was performed to augment the dataset. This technique introduces variability by flipping images along the vertical axis, which can help the model become invariant to horizontal orientations of the skin lesions. Such augmentation can prevent the model from overfitting to specific orientations in the training data[17].

Step 3: Resize image after padding and flipping

All images were resized to a specific resolution required by each model to ensure compatibility and consistent input dimensions during training. The original images before preprocessing ranged from 720 × 447 to 4024 × 6048 pixels, reflecting variability due to different devices and imaging conditions. This preprocessing step helped standardize the input format, facilitating batch processing and improving computational efficiency. The chosen

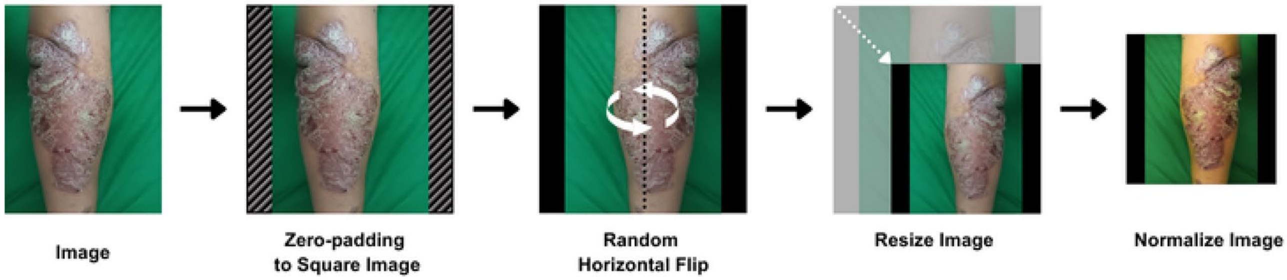

Fig. 2. Workflow of data preparation and augmentation: zero-padding, flip, resize, and normalize.

Disease	Public data	Siriraj Dermatology Databank	New collected image	Total
Training and validation set				
Psoriasis	73	80	1289	1442
Dermatophytosis	300	75	0	375
Eczema	428	83	0	511
Total	801	238	1289	2328
Testing set				
Psoriasis	18	21	321	360
Dermatophytosis	75	19	0	94
Eczema	107	21	0	128
Total	200	61	321	582
A pilot study				
Psoriasis	0	0	10	10
Dermatophytosis	8	2	0	10
Eczema	3	7	0	10
Total	11	9	10	30
Grand Total	1012	308	1620	2940

Table 1. Dataset overview.

resolution represented a trade-off between preserving essential visual features and maintaining a reasonable computational cost[18].

Step 4: Normalize image

The images were normalized to have a consistent mean and standard deviation. Normalization scales the pixel values to a standard range, which helps to accelerate convergence during training and improves numerical stability. This step ensures that the neural network or convolutional neural network treats all input features equally[19].

The final images for the development and testing of the algorithms had varying illumination effects and were divided into three sets: (i) training, (ii) validation, and (iii) testing data sets.

Implementation of Swin for skin lesion classification

An effective algorithm was developed based on CNN and Transformer architectures. The CNN models included eight existing architectures: AlexNet[20], DenseNet-121[21], EfficientNetV2[22], GoogLeNet[23], MobileNetV3[24], SqueezeNet[25], VGG-19[26], and ResNet-50[27]. Additionally, the eight Transformer-based models included ViT[28], Swin[29], CvT[30], DaViT[31], MaxViT[32], GC ViT[33], FastViT-S12[34], and SHViT-S1[35]. These architectures were trained and validated to classify each skin disease from skin images, and subsequently evaluated on a separate test set to confirm that their performance remained consistent. The parameters of both the CNN and Transformer models are shown in Table 2. To evaluate reasonable or unreasonable predictions by the architectures, Grad-CAM visualizations were generated to highlight which important regions of the image correspond to any decision of interest by an architecture.

The architecture of the Swin used for classifying skin lesion images was shown in Figure 3. Swin[29] is a hierarchical vision Transformer that introduces a novel shifted windowing mechanism for self-attention. It processes input images by dividing them into non-overlapping patches, which are then embedded into a sequence of tokens. The model consists of four stages, each comprising Swin Blocks that use either window-based multi-head self-attention (W-MSA) or shifted window-based self-attention (SW-MSA). These mechanisms allow the model to capture both local and global contextual information efficiently. Patch merging operations are applied between stages to progressively reduce spatial resolution and increase feature representation depth.

After feature extraction through the Swin blocks, the output is passed through an adaptive average pooling layer, which transforms variable-sized spatial features into a fixed-length feature vector. This vector is then input to a final fully connected layer that maps the features to class scores. In this study, we modified the final layer to contain three output nodes corresponding to the target classes: dermatophytosis, eczema, and psoriasis. The Swin is designed to be both computationally efficient and highly effective in capturing fine-grained image features relevant to medical imaging tasks such as skin lesion detection.

We used a pre-trained Swin model with weights from the ImageNet dataset[36]. To adapt it to our skin disease classification task, we applied transfer learning. This technique allowed a model that was trained on a large dataset, like ImageNet, to be reused for a different but related task. Instead of training from scratch, the model was fine-tuned on a smaller, specific dataset. This approach helped reduce training time, improves accuracy, and works well even when only a limited number of labeled medical images are available. Images of skin lesions were fed into the model to predict the corresponding disease class. To reduce overfitting and random split bias, the *k*-fold cross-validation technique was adopted, with the number of folds set to k = 5.

Method	Batch size	Loss function	Optimizer	Learning rate	Parameters	GFLOPs
AlexNet (2012)	4	Cross entropy loss	SGD	0.001	57.01 M	1.42
VGG19 (2014)	4	Cross entropy loss	SGD	0.001	139.58 M	39.28
GoogLeNet (2015)	4	Cross entropy loss	SGD	0.001	5.60 M	3.00
SqueezeNet (2016)	4	Cross entropy loss	SGD	0.001	0.73 M	1.47
ResNet-50 (2016)	4	Cross entropy loss	SGD	0.001	23.51 M	8.18
DenseNet-121 (2017)	4	Cross entropy loss	SGD	0.001	6.95 M	5.66
MobileNetV3 (2019)	4	Cross entropy loss	SGD	0.001	1.52 M	0.11
EfficientNetV2 (2021)	4	Cross entropy loss	SGD	0.001	20.18 M	5.70
ViT (2020)	4	Cross entropy loss	SGD	0.001	85.80 M	24.04
Swin (2021)	4	Cross entropy loss	SGD	0.001	86.74 M	21.10
CvT (2021)	4	Cross entropy loss	SGD	0.001	19.61 M	8.18
DaViT (2022)	4	Cross entropy loss	SGD	0.001	86.93 M	30.56
MaxViT (2022)	4	Cross entropy loss	SGD	0.001	30.40 M	10.96
GC ViT (2023)	4	Cross entropy loss	SGD	0.001	89.29 M	27.78
FastViT-S12 (2023)	4	Cross entropy loss	SGD	0.001	8.45 M	2.80
SHViT-S1 (2024)	4	Cross entropy loss	SGD	0.001	13.79 M	1.21

Table 2. Comparison of CNN and Transformer-based models in terms of parameter configuration, computational requirements, and structural complexity for skin disease image analysis. GFLOPs, Giga Floating Point Operations per second; M, million; SGD, Stochastic Gradient Descent.

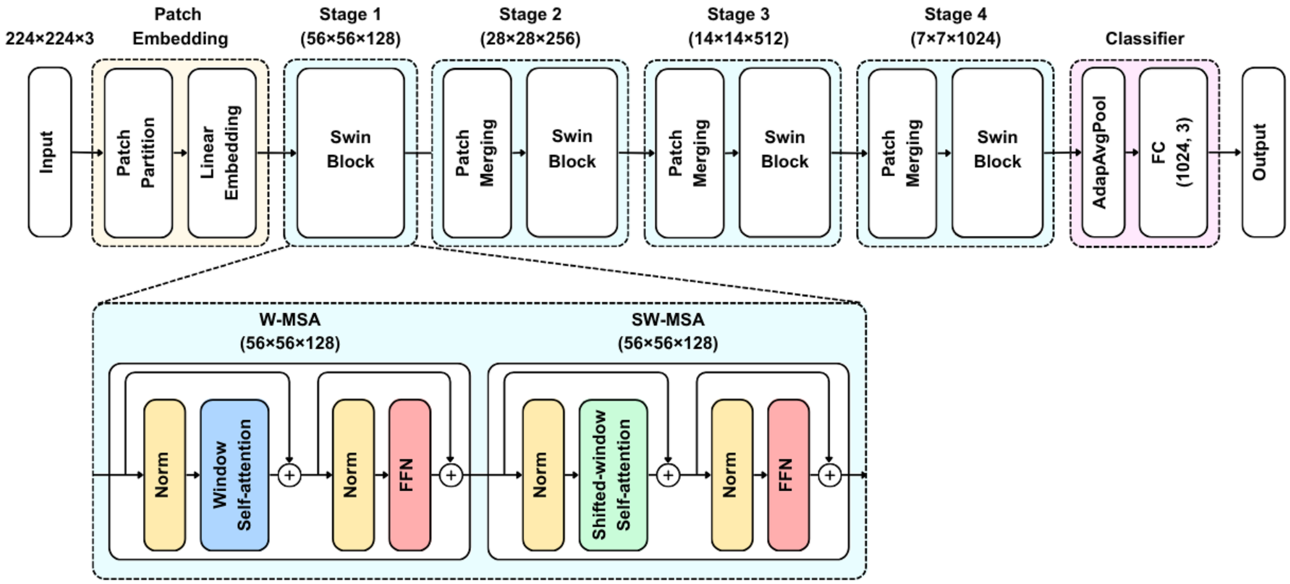

Fig. 3. Illustration of the Swin architecture, showing patch embedding, Swin blocks, and the classification pipeline.

Comparing the performance results between non-dermatologists and the deep learning model

To evaluate and compare the diagnostic accuracy of the best-performing CNN and Transformer models against clinicians without specialized dermatology training, a pilot study was conducted. For this purpose, 10 images per disease category, representing cases of psoriasis, eczema, and dermatophytosis, were selected. These images were not used during any training, validation, or testing processes to ensure unbiased evaluation. Participants were presented with the question, "What is the most likely diagnosis?" along with the options: "A. Psoriasis, B. Eczema, C. Dermatophytosis." The images were randomly integrated into an online questionnaire created using Google Forms.

The set of 30 images (10 per diagnostic class) used in the human-AI comparison was selected based on equal class representation and diagnostic clarity, rather than formal sample size calculation. This design aimed to facilitate a focused pilot comparison rather than a fully powered inferential study. Each image represented a distinct case from the test set, ensuring no overlap with the training or validation data. While the sample enabled qualitative and quantitative benchmarking across physician groups and AI, we acknowledge that it may be underpowered for detecting smaller inter-group differences and should be interpreted as exploratory.

Thirty clinicians without specialized dermatology training were recruited to voluntarily provide diagnoses for the 30 images. Participants were divided into two groups based on their clinical experience: the intern group, consisted of medical interns in their first postgraduate year of clinical training. These individuals had recently graduated from medical school and were working under supervision in hospital-based settings as part of their national service. And the internist group (n = 15) included board-certified internal medicine physicians with at least three years of clinical experience. They were involved in both outpatient and inpatient care at secondary hospitals but had no formal dermatology fellowship or extended dermatology rotations.

This comparison was designed to capture the influence of clinical experience on diagnostic performance among non-dermatologists. Interns represent entry-level clinical exposure, while internists embody more seasoned generalists. By including both, we aimed to assess how AI performance compares across different experience levels typical in primary and secondary care settings, where dermatology expertise is often limited.

Statistical analysis

Performance was evaluated using four metrics: precision, recall, F1 score, and accuracy, each ranging from 0 (very poor) to 1 (perfect). where TP = True Positives, FP = False Positives, TN = True Negatives, and FN = False Negatives. The formulas and interpretations of these metrics are provided below.

Precision is the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

where TP is True Positives and FP is False Positives. Precision is critical when the cost of false positives is high. For example, in medical diagnosis, predicting a disease when it is not present can lead to unnecessary treatment and anxiety. Precision helps ensure that when the model predicts a positive class, it is very likely to be correct. High precision indicates that the model has a low false positive rate, which is crucial for applications where false positives can be particularly costly.

Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where FN is False Negatives. Recall is important when the cost of false negatives is high. For instance, in the same medical diagnosis example, missing a disease (false negative) can be very dangerous. Recall ensures that the model identifies as many actual positives as possible. High recall means that the model has a low false negative rate, which is essential in applications where missing a positive case could have serious consequences[37].

F1 Score is the harmonic mean of Precision and Recall. It provides a single metric that balances both concerns.

$$\text{F1 Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{3}$$

The F1 Score is useful when you need to find a balance between Precision and Recall. It is particularly valuable when the classes are imbalanced and one class is significantly rarer than the other. It provides a single measure that accounts for both false positives and false negatives, making it a good indicator of the model's overall effectiveness in identifying positive instances without being biased by the majority class[38].

Accuracy is the ratio of correctly predicted observations to total observations.

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

where TN is True Negatives. Accuracy is a straightforward metric that provides an overall view of the model's performance. However, it can be misleading in cases of class imbalance. For example, if 95% of the data belongs to one class, a model that predicts the majority class all the time will have high accuracy but poor performance in terms of Precision, Recall, and F1 Score for the minority class. Thus, while accuracy gives a general performance measure, it should be considered alongside the other metrics to get a full picture of model performance[39].

Using these four metrics together provides a comprehensive evaluation of the model's performance, ensuring it performs well not only overall but also across different aspects of classification performance.

To assess human diagnostic performance, evaluators' responses were analyzed to calculate true positives, false positives, true negatives, and false negatives for each diagnostic category. Confusion matrices were constructed to visualize the distribution of predictions across actual categories, providing insights into patterns of correct and incorrect diagnoses. The matrices highlighted areas where human evaluators struggled most, particularly when differentiating eczema from psoriasis or distinguishing dermatophytosis from other conditions. Statistical p-values were calculated using two-proportion tests to evaluate whether differences in diagnostic accuracy between evaluator groups—including interns and internists—were significant.

In parallel, the diagnostic performance of the deep learning model was evaluated using the same test dataset. A confusion matrix was generated to compare its predictions with the ground truth labels. The deep learning model's confusion matrix was directly compared with those of interns and internists to identify areas where the AI showed superior or inferior performance relative to human evaluators. Key performance metrics including overall accuracy, precision, recall, and F1 score were derived from all confusion matrices. One-proportion tests were applied to assess statistically significant differences between AI and human performance across these key diagnostic metrics.

All statistical analyses were conducted using MedCalc Statistical Software, version 22.013 (MedCalc Software Ltd., Ostend, Belgium) and IBM Statistical Package for the Social Sciences (SPSS) Statistics, version 26 (IBM Corp., Armonk, NY, USA).

Results

Performance of deep learning models in diagnosing skin lesions from images

To identify the CNN and Transformer architectures that performs best for skin lesion classification in skin images. Table 2 presented the main parameters of sixteen models, which were considered during the selection process. Batch Size represents the number of training examples used in a single iteration. The Loss Function measures the error in the model's predictions. The Optimizer is the algorithm used to update the model's weights. Learning Rate indicates the step size during each iteration, as the model seeks to minimize the loss function. Parameters refer to the total number of learnable parameters in the model. Finally, Giga Floating Point Operations per second (GFLOPs) reflect the computational complexity of the model. All models in the comparison use a batch size of 4, ensuring a fair and consistent training process. They all employ the cross-entropy loss function and the stochastic gradient descent optimizer with a learning rate of 0.001. This consistency allows for a direct comparison of their inherent capabilities without variability in the training setup. We used the k-fold cross-validation technique to avoid overfitting and random split bias. We set *k* to 5.

Table 3 presents the performance comparison of CNN and Transformer-based architectures on both training and validation datasets for the classification of dermatophytosis, eczema, and psoriasis. The table includes precision, recall, F1-score, and accuracy values for each class, as well as the overall classification performance of each model. To make the results easier to follow, the models were grouped into three categories based on validation accuracy: high-performing models (accuracy$\geq$0.900), moderate-performing models (accuracy between 0.800 and 0.899), and low-performing models (accuracy<0.800). The high-performance group comprised ten models—VGG19, SqueezeNet, ResNet-50, MobileNetV3, EfficientNetV2, ViT, Swin, DaViT, MaxViT, and GC ViT—all of which achieved an accuracy of 0.900 or higher. This group included both advanced CNNs and most Transformer-based architectures. Their strong performance highlights the ability of these models to capture complex visual features from macroscopic skin images. Among them, Swin and ViT achieved the highest accuracy scores, demonstrating the growing effectiveness of Transformer-based models in medical image classification. Specifically, Swin and ViT achieved F1 scores above 0.82 for dermatophytosis, 0.87 for eczema, and 0.97 for psoriasis, reflecting their consistent and robust performance across all target classes. The moderate-performance group included five models: AlexNet, GoogLeNet, DenseNet-121, CvT, and FastViT-S12. With accuracies ranging from 0.800 to 0.899, these models achieved acceptable performance but fell short of the top-performing architectures. Although relatively efficient in terms of computational cost, they may have limited ability to capture the subtle and complex visual patterns necessary for distinguishing between clinically similar conditions, particularly dermatophytosis and eczema. The low-performance group contained only one model,

	Dermatophytosis			Eczema			Psoriasis			All classes			
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy
AlexNet	0.731	0.741	0.735	0.815	0.796	0.805	0.968	0.972	0.970	0.838	0.836	0.837	0.896
VGG19	0.813	0.759	0.783	0.843	0.877	0.858	0.981	0.982	0.982	0.879	0.873	0.874	0.923
GoogLeNet	0.686	0.919	0.783	0.807	0.761	0.773	0.991	0.898	0.937	0.828	0.859	0.831	0.871
SqueezeNet	0.707	0.811	0.755	0.850	0.773	0.809	0.978	0.972	0.975	0.845	0.852	0.846	0.902
ResNet-50	0.731	0.903	0.806	0.934	0.810	0.867	0.987	0.972	0.979	0.884	0.895	0.884	0.925
DenseNet-121	0.585	0.911	0.691	0.825	0.789	0.794	0.987	0.792	0.857	0.799	0.831	0.781	0.810
MobileNetV3	0.807	0.770	0.776	0.836	0.847	0.834	0.980	0.974	0.977	0.874	0.864	0.862	0.914
EfficientNetV2	0.731	0.922	0.809	0.892	0.861	0.874	0.994	0.929	0.959	0.872	0.904	0.881	0.913
ViT	0.835	0.814	0.822	0.863	0.879	0.870	0.976	0.974	0.975	0.891	0.889	0.889	0.928
Swin	0.845	0.819	0.831	0.882	0.894	0.888	0.983	0.985	0.984	0.904	0.899	0.901	0.938
CvT	0.714	0.532	0.576	0.694	0.798	0.736	0.954	0.951	0.952	0.787	0.761	0.755	0.851
DaViT	0.799	0.757	0.774	0.836	0.861	0.847	0.979	0.979	0.979	0.871	0.866	0.867	0.918
MaxViT	0.798	0.749	0.766	0.813	0.871	0.838	0.977	0.963	0.970	0.863	0.861	0.858	0.909
GC ViT	0.800	0.786	0.791	0.859	0.847	0.852	0.975	0.981	0.978	0.878	0.872	0.874	0.921
FastViT-S12	0.704	0.692	0.696	0.749	0.814	0.779	0.981	0.953	0.967	0.811	0.820	0.814	0.881
SHViT-S1	0.582	0.254	0.333	0.556	0.769	0.642	0.915	0.913	0.914	0.684	0.646	0.630	0.777

Table 3. Performance comparison of CNN and Transformer-based architectures on the training and validation datasets for classifying dermatophytosis, eczema, and psoriasis, along with overall classification performance. Results in this table represent internal validation performance only. The results are shown as average values. *CNN*, Convolutional Neural Network.

SHViT-S1, which achieved an accuracy of 0.777. Its relatively poor performance suggests that the model's design may not be well suited to this classification task.

Table 4 presents the test set performance of CNN and Transformer-based models, confirming the trends observed during validation (Table 3). Overall, most models demonstrated consistent generalization, with top-performing architectures retaining high accuracy and F1 scores across all classes. Swin and ViT, which achieved the highest validation accuracy in Table 3, remained the best-performing models on the test set, both achieving perfect or near-perfect scores for psoriasis (F1 = 1.000) and high F1 scores for dermatophytosis and eczema. Their overall test accuracies were 0.967 and 0.960, respectively, demonstrating stable performance when applied to unseen data.

Among the evaluated models, the Swin has a parameter count of 86.74 million and requires 21.10 GFLOPs, placing it among the more resource-intensive models in the comparison. When compared with other Transformer-based models, Swin demonstrates a balanced trade-off between model size and computational demand. For instance, DaViT and GC ViT have slightly higher parameter counts (86.93 M and 89.29 M, respectively), yet require more computation (30.56 and 27.78 GFLOPs). Similarly, ViT has a comparable parameter count (85.80 M) but consumes more computational resources (24.04 GFLOPs), suggesting that Swin is relatively more efficient in terms of design. On the other hand, lightweight Transformers such as CvT, MaxViT, FastViT-S12, and SHViT-S1 operate with fewer than 20 million parameters and under 10 GFLOPs, making them more suitable for environments with limited computational capacity—though typically with some compromise in performance. Compared to CNN models, Swin requires higher resource consumption than compact architectures like SqueezeNet (0.73 M, 1.47 GFLOPs), MobileNetV3 (1.52 M, 0.11 GFLOPs), and GoogLeNet (5.60 M, 3.00 GFLOPs), but this is offset by its stronger classification performance in complex tasks such as skin disease recognition.

Figure 4 presents confusion matrices for all CNN and Transformer-based models evaluated on the test dataset for three-class skin condition classification. The color intensity in each cell represents the proportion of predictions, with darker shades indicating higher frequencies. These matrices visualize both correct predictions (diagonal elements) and misclassification patterns, providing insights into each model's classification behavior. The visualizations were generated using predictions from the best-performing fold, representing the highest-performing results of each model during evaluation.

The analysis reveals notable performance differences across architectures. The Swin shows a perfectly diagonal matrix, meaning all test cases were correctly classified with no misidentifications. This suggests Swin is highly effective at distinguishing between challenging skin conditions—particularly between dermatophytosis and eczema, which are frequently misclassified by both deep learning models and human observers. In contrast, several CNN architectures exhibited varying degrees of classification confusion. AlexNet and GoogLeNet showed substantial misclassification between dermatophytosis and eczema categories, while more recent architectures like EfficientNetV2, MobileNetV3, and ResNet-50 demonstrated improved but imperfect performance with occasional classification errors. Among other Transformer models, ViT, DaViT, and GC ViT generally performed well but still showed some confusion between eczema and dermatophytosis classes. Notably, only the Swin achieved perfect separation across all three diagnostic categories, highlighting its superior ability to capture subtle visual distinctions in clinical skin images. As illustrated in Figure 4, the Swin achieved perfect diagonal separation across all diagnostic classes, with no misclassifications observed. By contrast, several CNN

	Dermatophytosis			Eczema			Psoriasis			All classes			
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy
AlexNet	0.813	0.840	0.822	0.896	0.680	0.765	0.836	1.000	0.910	0.848	0.840	0.832	0.840
VGG19	0.833	0.820	0.821	0.887	0.820	0.846	0.927	1.000	0.962	0.883	0.880	0.876	0.880
GoogLeNet	0.810	1.000	0.894	0.945	0.760	0.836	1.000	0.940	0.965	0.918	0.900	0.898	0.900
SqueezeNet	0.842	0.820	0.827	0.861	0.820	0.834	0.927	0.980	0.951	0.877	0.873	0.871	0.873
ResNet-50	0.808	1.000	0.893	1.000	0.760	0.863	1.000	1.000	1.000	0.936	0.920	0.919	0.920
DenseNet-121	0.802	1.000	0.888	0.907	0.800	0.838	1.000	0.800	0.857	0.903	0.867	0.861	0.867
MobileNetV3	0.881	0.880	0.879	0.863	0.860	0.860	0.980	0.980	0.980	0.908	0.907	0.906	0.907
EfficientNetV2	0.821	1.000	0.901	1.000	0.800	0.888	1.000	0.980	0.989	0.940	0.927	0.926	0.927
ViT	0.897	1.000	0.945	1.000	0.880	0.935	1.000	1.000	1.000	0.966	0.960	0.960	0.960
Swin	0.925	0.980	0.951	0.980	0.920	0.948	1.000	1.000	1.000	0.968	0.967	0.967	0.967
CvT	0.795	0.640	0.692	0.789	0.600	0.661	0.742	1.000	0.847	0.775	0.747	0.733	0.747
DaViT	0.962	0.900	0.927	0.947	0.900	0.919	0.917	1.000	0.955	0.942	0.933	0.934	0.933
MaxViT	0.879	0.820	0.844	0.881	0.860	0.869	0.930	1.000	0.963	0.897	0.893	0.892	0.893
GC ViT	0.900	0.840	0.867	0.906	0.900	0.898	0.948	1.000	0.972	0.918	0.913	0.913	0.913
FastViT-S12	0.874	0.720	0.783	0.832	0.880	0.851	0.895	0.980	0.935	0.867	0.860	0.857	0.860
SHViT-S1	0.510	0.160	0.229	0.573	0.700	0.619	0.673	1.000	0.803	0.585	0.620	0.550	0.620

Table 4. Performance comparison of CNN and transformer-based architectures on the independent test dataset for classifying dermatophytosis, eczema, and psoriasis, along with overall classification performance. These results confirm generalizability on unseen data. The results are shown as average values. *CNN*, Convolutional Neural Network.

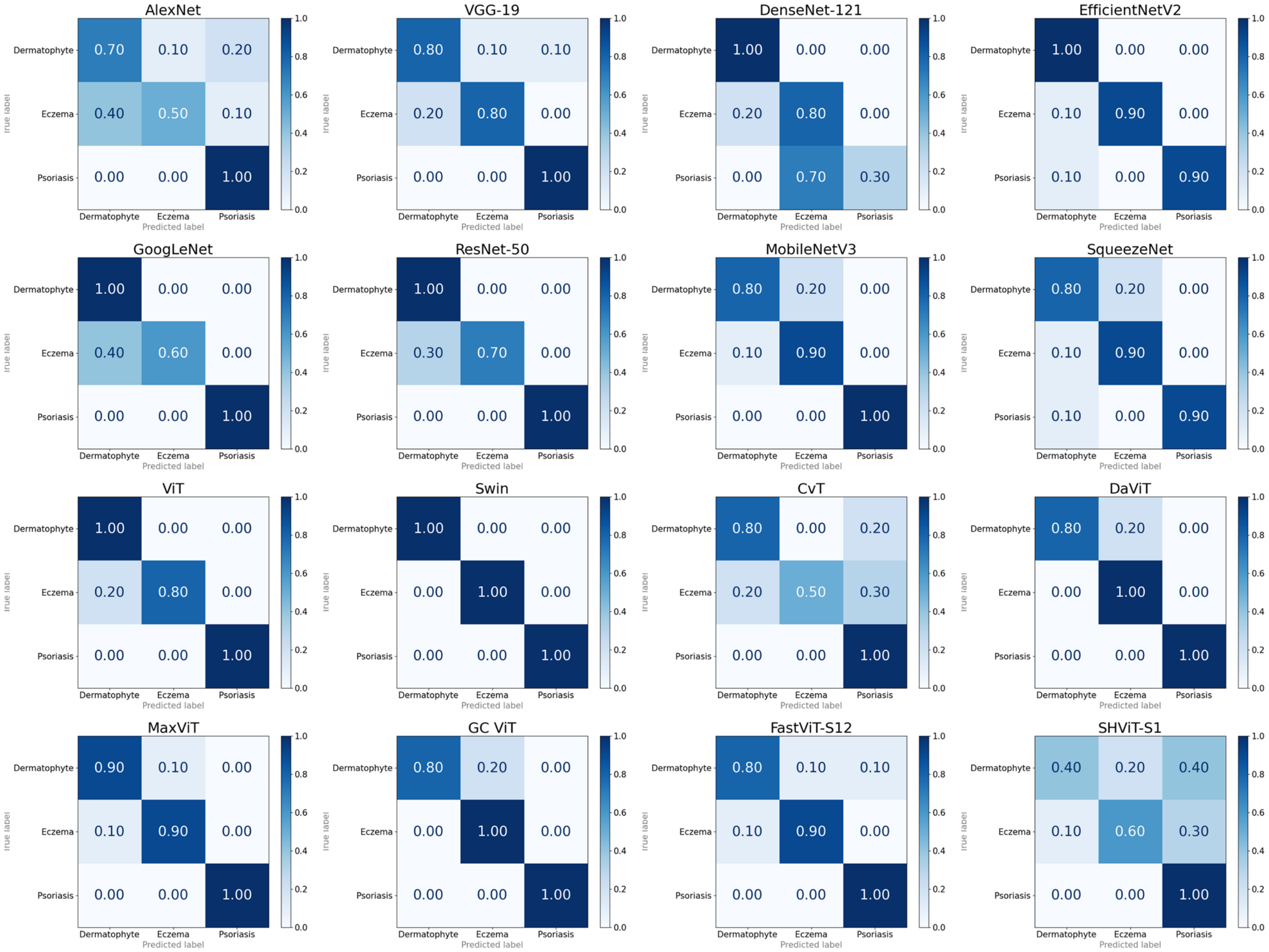

Fig. 4. Confusion matrices for CNN and Transformer models on three-class skin condition classification.

and Transformer models demonstrated residual confusion, particularly between eczema and dermatophytosis. To complement these results, Figure 11 depicts the confusion matrix of physician responses, showing frequent misclassification of eczema as psoriasis or dermatophytosis. Taken together with Figure 4, which demonstrates perfect diagonal separation by the Swin, these findings underscore the contrast between AI and human evaluators, while the model consistently achieved flawless classification across all categories, physicians which include both interns and internists struggled most with eczema, often confusing it with other erythematous conditions. This pictorial comparison highlights not only the superior accuracy of the Swin but also its potential to address recurrent diagnostic blind spots in clinical practice. This visual comparison underscores the robustness of the Swin relative to both other architectures and human evaluators. Although Swin and ViT have comparable parameter counts (86.74M vs. 85.80M, respectively), Swin consistently outperformed ViT across all evaluation metrics (Tables 3, 4). This performance gap can be attributed to Swin's architectural innovations, particularly its shifted window self-attention mechanism and hierarchical feature representation. While ViT applies global self-attention at a single resolution, Swin introduces local window-based attention with overlapping regions through shifting windows, which enables efficient modeling of both local and long-range dependencies. Furthermore, Swin's hierarchical design—progressively merging patches across layers—allows the model to capture multiscale contextual features that are crucial in dermatological imaging, where lesions often exhibit fine-grained textures and spatial variability. This design not only improves computational efficiency but also enhances the model's ability to focus on clinically meaningful regions, leading to improved classification performance across all skin disease categories in our study. This result supports previous findings (Tables 3, 4), where Swin achieved the highest F1 scores and accuracy, further confirming its robustness and precision in clinical image-based skin disease classification.

As mentioned above, we can conclude that the Swin consistently outperformed all other models across multiple evaluation metrics. It achieved the highest test accuracy of 0.967 with strong F1 scores for all three conditions (Tables 3, 4). Most importantly, Figure 4 shows that Swin was the only model to achieve perfect classification with zero misclassifications, successfully distinguishing between the visually similar dermatophytosis and eczema conditions that challenge other models. While Swin requires moderate computational resources (86.74M parameters, 21.10 GFLOPs), it is more efficient than comparable Transformers like ViT and DaViT (Table 2). This demonstrates Swin's suitability for skin lesion classification tasks.

Clinical interpretation and evaluation of Grad-CAM outputs

Purpose and interpretability of Grad-CAM

Grad-CAM is a widely used method for visualising which regions of an input image influence a model's prediction. In medical applications—where accountability and transparency are critical—such spatial attention visualisations offer an interpretability layer that bridges deep learning outputs with clinician intuition. By tracing the internal reasoning process of convolutional and Transformer-based models, Grad-CAM provides insight not just into *what* was predicted, but *why*. In dermatology, where lesion localisation is diagnostic, Grad-CAM is particularly valuable in assessing whether a model's prediction is rooted in clinically meaningful regions. This contextual interpretability is crucial for building trust in AI-assisted skin disease classification.

Observed attention patterns across model families

Figures 5, 6, 7, 8, 9 and 10 display Grad-CAM visualisations for 16 models (eight CNN-based and eight Transformer-based) across three skin disease categories: dermatophytosis, eczema, and psoriasis. These heatmaps show distinct patterns of spatial attention between architectural families. Among CNN-based models, architectures such as AlexNet, GoogLeNet, DenseNet-121, and SqueezeNet were more prone to misclassification. Their Grad-CAM visualizations often revealed attention directed toward irrelevant background or unaffected skin regions—especially in conditions like dermatophytosis and eczema—suggesting a tendency to rely on artefactual or contextual cues. Correct classifications in CNNs were generally associated with narrowly focused red highlights on a portion of the lesion, whereas incorrect predictions frequently occurred when lesion areas were neglected or highlighted in cooler colours. By contrast, Transformer-based models generally demonstrated stronger alignment with clinically meaningful regions. Swin and ViT, in particular, consistently generated accurate predictions with heatmaps focused squarely on lesion zones. Grad-CAM visualizations with clinical annotations further illustrate this alignment, highlighting hallmark diagnostic cues such as silvery scales in psoriasis, diffuse erythema in eczema, and the raised peripheral rim with central clearing in dermatophytosis. These annotated comparisons reinforce that the highlighted regions correspond to features routinely used by clinicians in diagnostic reasoning. Swin maintained precise attention on pathologic areas across all three disease categories. Other Transformer models—such as DaViT, MaxViT, and GC ViT—also showed reliable localisation but occasionally included surrounding non-lesional skin within their focus. Across the board, correct predictions were associated with red-highlighted lesion centres, while failures typically involved dispersed or misplaced focus. These findings underscore the critical role of spatial attention in dermatologic AI classification.

To move beyond architectural trends, we further examined how these attention patterns aligned with clinical reasoning through structured expert review.

Clinical review of Grad-CAM outputs

To assess the clinical plausibility of the Grad-CAM visualisations, two board-certified dermatologists (C.W. and C.C.) jointly reviewed a representative subset of Grad-CAM heat-maps through a structured discussion. The reviewers assessed whether attention maps corresponded to key diagnostic features for psoriasis, eczema, and dermatophytosis, and whether the visual patterns aligned with real-world clinical reasoning.

Overall, CNN models tended to produce narrower, more localized attention focused on high-contrast features such as silvery scales or annular borders. These maps were generally easier to interpret but occasionally failed to capture the full lesion extent, particularly for diffuse conditions like eczema. In contrast, Transformer models demonstrated broader spatial coverage and were more likely to capture composite lesion patterns, such as both peripheral rim and central clearing in dermatophytosis. However, this breadth sometimes came at the expense of specificity, with occasional attention spillover into non-lesional skin or background artefacts.

The reviewers concluded that while CNN models are often more intuitive, Transformer models better mimic holistic diagnostic strategies used in practice. A structured summary of these observations across all disease categories and model types is provided in Table 5.

Performance of novices and experienced non-dermatologists in diagnosing skin lesions from images

Among non-dermatologist physicians, internists demonstrated moderately higher diagnostic performance than interns across all disease categories (see Supplementary Table S1, Figure 11). While both groups showed similar trends in misclassification patterns, internists achieved higher recall and F1 scores in eczema and dermatophytosis, likely reflecting their broader clinical experience.

Comparing the performance results between nondermatologists and the deep learning model in classifying the skin diseases

Table 6 presents a performance comparison between 30 physicians (comprising interns and internists) and the Swin model in diagnosing dermatophytosis, eczema, and psoriasis from clinical images. Across all disease categories, Swin consistently outperformed human evaluators in precisions, recall and F1 score. Differences between AI and human performance were statistically significant for nearly all metrics (all $p < 0.001$), except for dermatophytosis precision ($p = 0.467$). Notably, the largest performance gap was observed in recall for psoriasis (Swin: 1.000 vs. physicians: 0.716; $p < 0.001$; 95% CI: 0.722 to 0.974).

For dermatophytosis, the Swin outperformed physicians across all evaluation metrics, achieving a recall (0.980), precision (0.925), and F1 score (0.951). In comparison, physicians attained a recall (0.737), precision (0.890), and F1 score (0.774).

In the classification of eczema, the Swin demonstrated superior performance, achieving higher precision (0.980), recall (0.920), and F1 score (0.948). In contrast, physicians recorded lower scores across all metrics, with a precision of 0.864, recall of 0.694, and F1 score of 0.843.

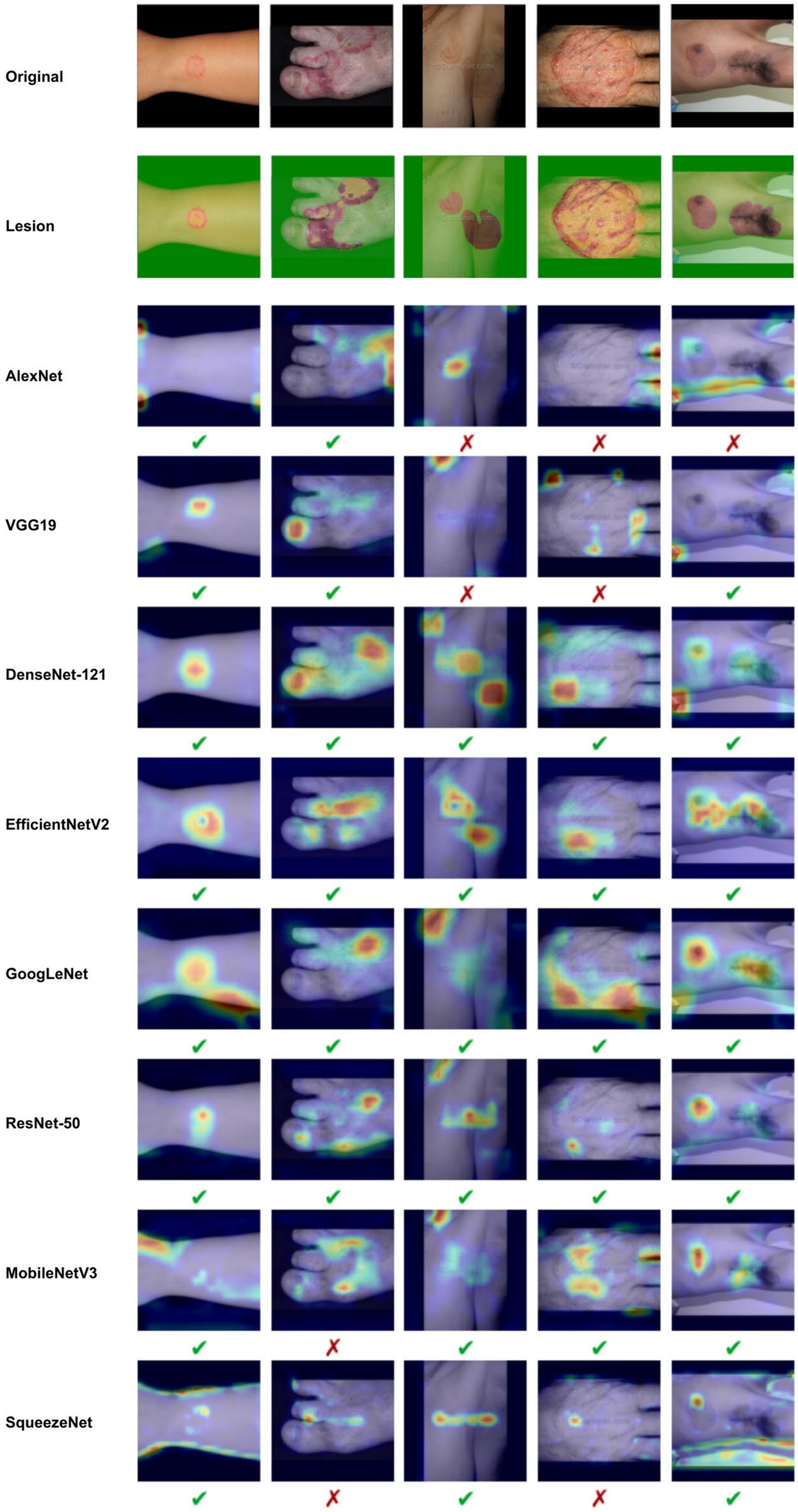

Fig. 5. Comparative Grad-CAM visualizations across CNN models for dermatophytosis case, illustrating focus areas associated with model decisions.

For psoriasis, Swin obtained perfect scores in all three metrics including precision, recall, and F1 score (1.000 each). Physicians demonstrated lower performance, with a precision of 0.877, recall of 0.716, and F1 score of 0.807.

When evaluating overall classification performance across all classes, Swin again outperformed human physicians, achieving a precision of 0.968, recall of 0.967, F1 score of 0.967, and overall accuracy of 0.967. In comparison, physicians achieved 0.801 precision, 0.800 recall, 0.800 F1 score, and 0.867 accuracy.

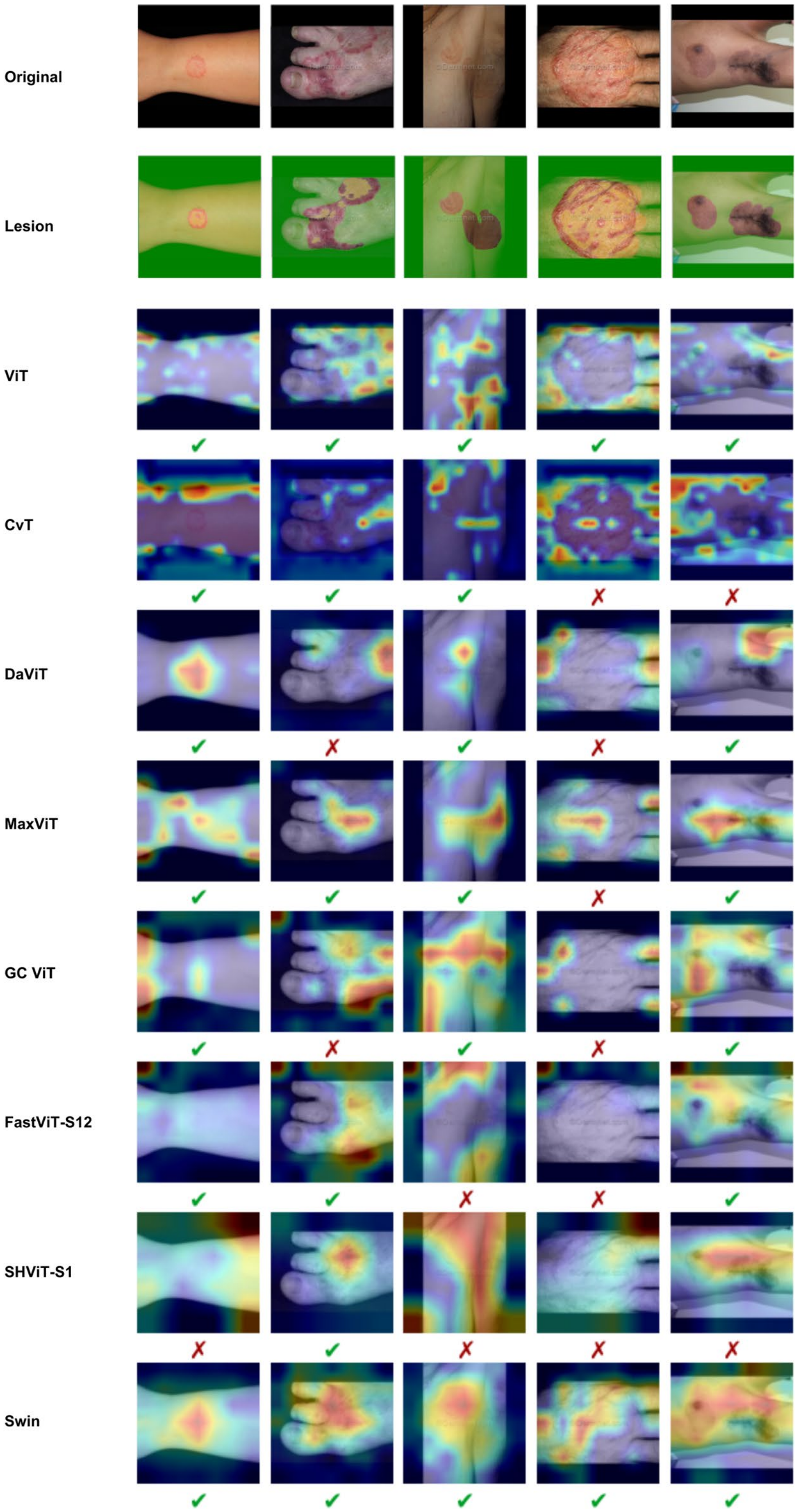

Fig. 6. Comparative Grad-CAM visualizations across Transformer-based models for dermatophytosis case, illustrating focus areas associated with model decisions.

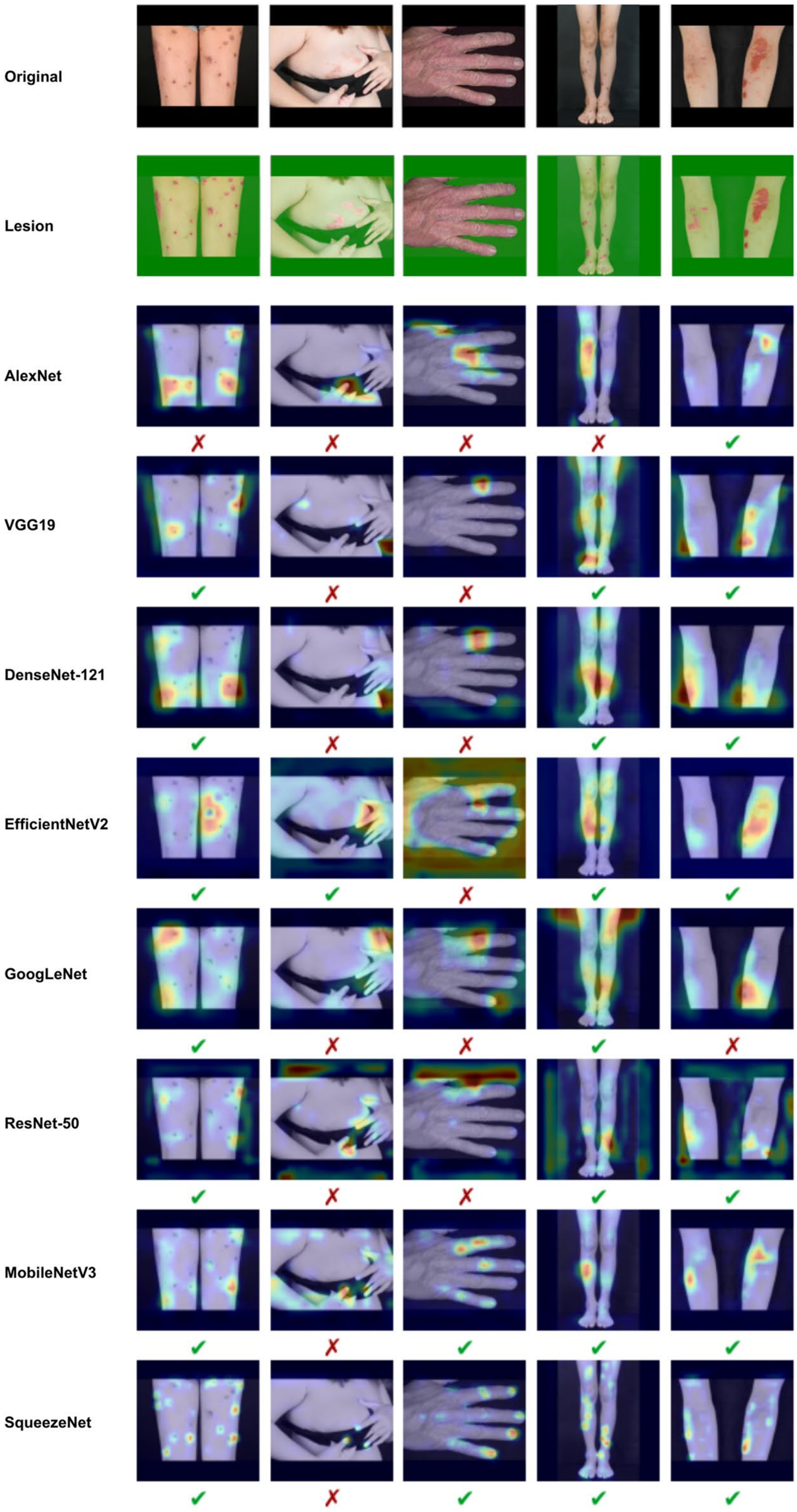

Fig. 7. Comparative Grad-CAM visualizations across CNN models for eczema case, illustrating focus areas associated with model decisions.

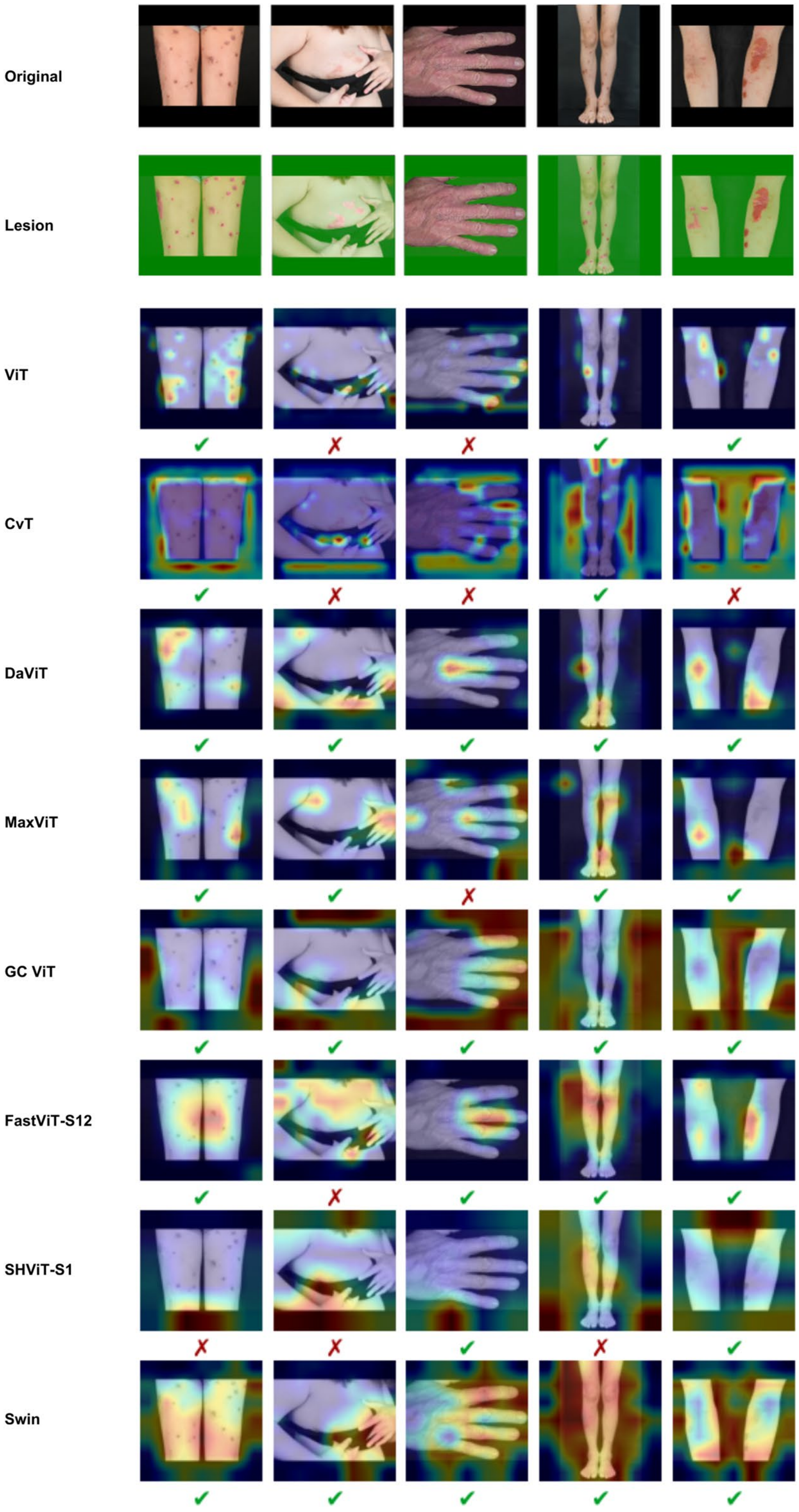

Fig. 8. Comparative Grad-CAM visualizations across Transformer-based models for eczema case, illustrating focus areas associated with model decisions.

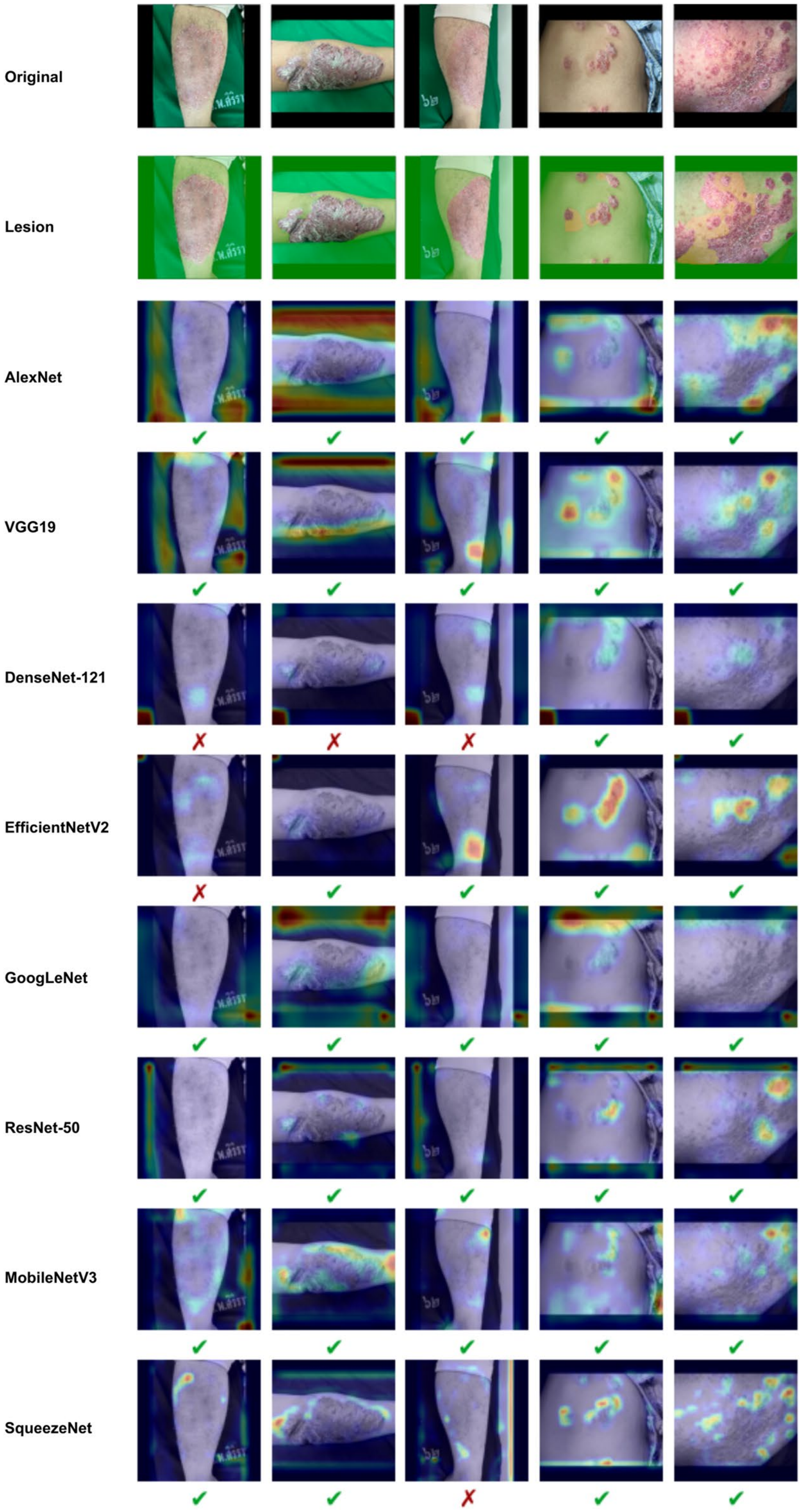

Fig. 9. Comparative Grad-CAM visualizations across CNN models for psoriasis case, illustrating focus areas associated with model decisions.

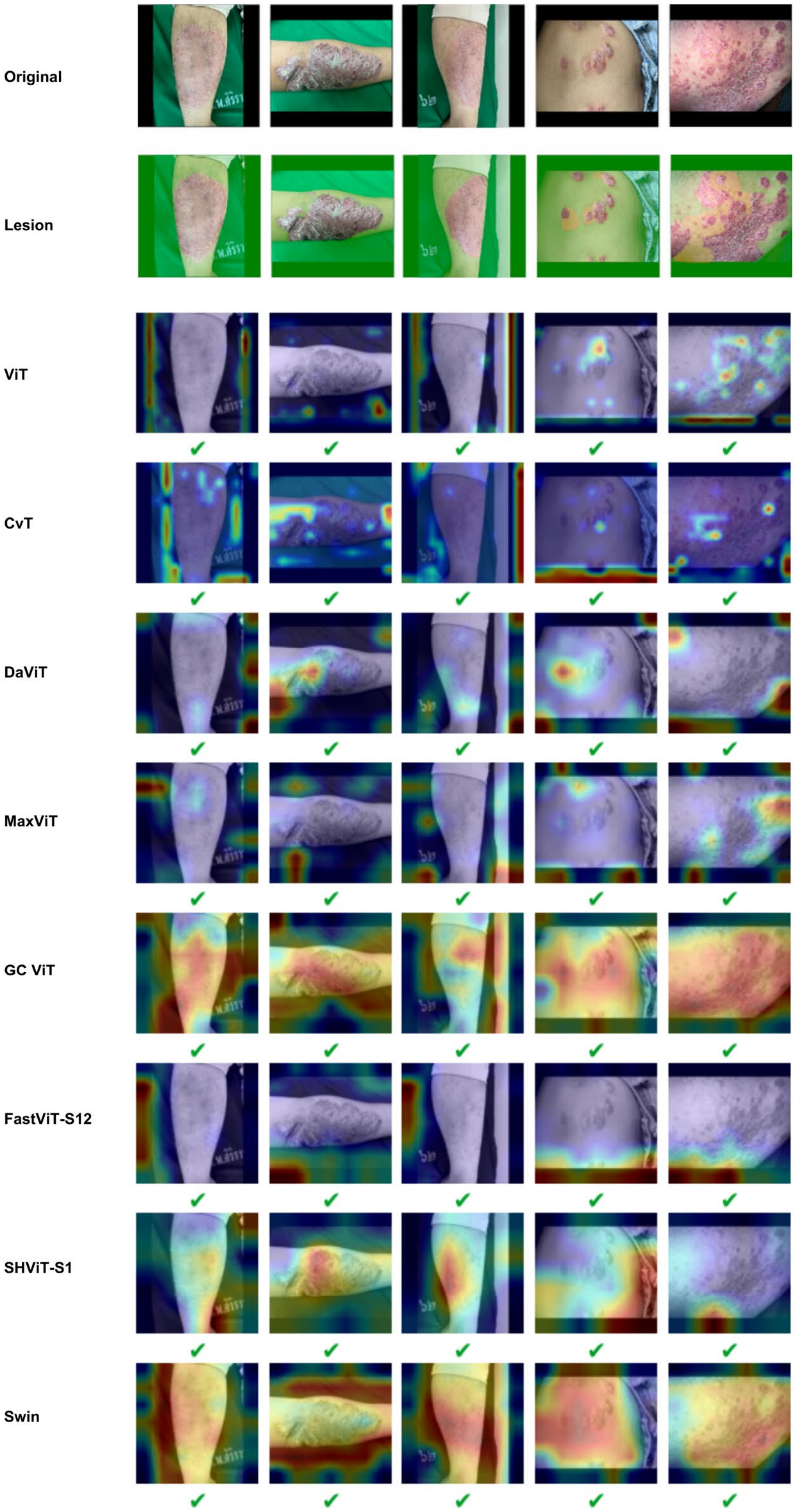

Fig. 10. Comparative Grad-CAM visualizations across Transformer-based models for psoriasis case, illustrating focus areas associated with model decisions.

Disease	CNN models	Transformer models	Expert summary
Psoriasis	Focused on silvery scale; localized attention on hyperkeratotic zones; occasionally incomplete	Outlined full plaque area; broader focus including shadows or non-lesional skin	Both model types captured key plaque features; CNN more focal, Transformer more comprehensive
Eczema	Captured focal inflammation but often missed peripheral cues	Highlighted diffuse erythema and excoriations; better coverage of flexural skin	Transformers better at capturing diffuse patterns; CNN limited in scope
Dermatophytosis	Focused on segment of annular border; strong local contrast recognition	Highlighted both peripheral rim and central clearing	Transformers mimicked expert reasoning more closely; CNNs more selective

Table 5. Expert summary of Grad-CAM attention patterns across model types and skin diseases.

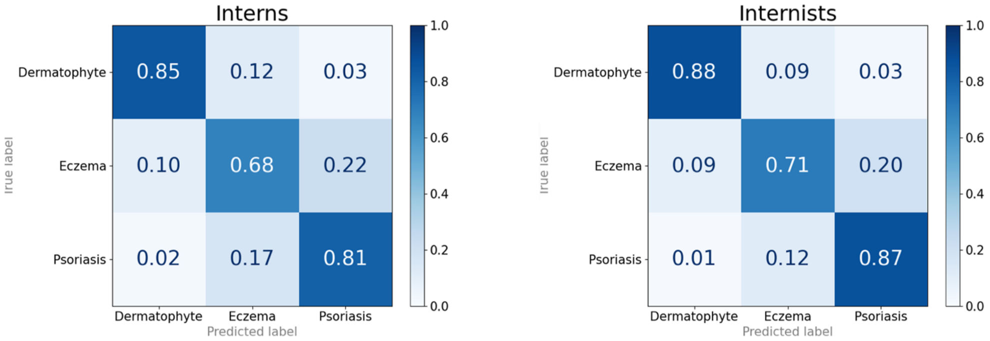

Fig. 11. Comparative confusion matrices for interns and internists in diagnosing three skin conditions.

Discussion

The results of this study underscore the considerable promise of transformer-based architectures, particularly the hierarchical Swin, as diagnostic tools for dermatological conditions such as psoriasis, dermatophytosis, and eczema. In our experiments, Swin consistently outperformed every comparator model and the physician cohort across precision, recall, and F1 metrics. These findings corroborate a growing body of evidence: Mohan et al. reported that a Swin-based pipeline achieved a macro-F1 of 0.95 and 93 % accuracy on a 31-class skin-disease dataset, clearly surpassing CNN baselines[40], a recent systematic review likewise concluded that vision-transformer families set the current state of the art in cutaneous-image recognition[41] and a conference study that applied Swin to 24 skin conditions documented$>$ 5 % accuracy gains over ResNet-50 and EfficientNet benchmarks[42]. Collectively, these lines of evidence position hierarchical vision transformers as leading candidates for real-world dermatological decision support and justify further prospective clinical validation.

Beyond architecture design, the integration of meta-heuristic optimization algorithms with deep learning models has shown significant promise in a variety of medical-imaging tasks. Saber et al.[43] employed a hybrid ensemble framework that combined deep networks with meta-heuristic algorithms for breast-tumor classification, achieving notable improvements in diagnostic accuracy. Elbedwehy et al.[44] likewise incorporated advanced optimization strategies with neural networks to enhance kidney-disease detection, underscoring the value of feature selection and hyper-parameter tuning. Khaled et al.[45] further demonstrated that coupling adaptive CNNs with the grey-wolf optimizer boosted breast-cancer diagnostic performance. Taken together, these studies suggest that optimization-driven enhancements could further strengthen models like Swin in future dermatologic applications.

Moreover, Swin also outperformed non-dermatologists which are interns and internists, across key metrics like diagnostic precision, recall, F1 scores, and overall accuracy. This performance demonstrates its ability to bridge gaps in clinical expertise, particularly for complex and variable conditions like eczema.

An important characteristic of advanced Transformer and CNN models, such as Swin, can be observed through the use of Grad-CAM, which provides insights into the regions of an image that the model prioritizes for its predictions. Annotated Grad-CAM visualizations explicitly mark clinically meaningful features as confirmed by expert dermatologists, demonstrating that the model's attention is not arbitrary but grounded in features fundamental to diagnosis. This alignment with clinical reasoning strengthens clinician trust and supports integration of such models into medical education and teledermatology workflows. Apart from accurately identifying lesioned areas, the model's predictions align closely with clinical reasoning principles taught in medical education. For instance, in cases of dermatophytosis, clinical training emphasizes recognizing a central clearing with an active, raised border. Grad-CAM visualizations from Swin effectively focus on these diagnostic features. For eczema, the attention maps highlight diffuse and inflamed patches, consistent with the condition's varied presentations and clinical complexity. For psoriasis, the Grad-CAM emphasize well-demarcated plaques with scaling, features that are central to its clinical identification. These observed attentions mimicked the reasoning patterns used by human experts. The insights provide a clear and interpretable basis for the model's predictions while ensuring consistency with established clinical practices.

The diagnostic superiority of Swin can be attributed to its extensive and diverse training dataset of 2,940 images, enabling it to capture subtle distinctions between dermatological conditions. While the combined

Method	Dermatophytosis			Eczema			Psoriasis			All classes			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy
Physicians	0.890	0.737	0.774	0.864	0.694	0.843	0.877	0.716	0.807	0.801	0.800	0.800	0.867
Swin	0.925	0.980	0.951	0.980	0.920	0.948	1.000	1.000	1.000	0.968	0.967	0.967	0.967
p value	0.467	<0.001	<0.001	<0.001	<0.001	0.010	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.002
95% CI interval	0.722 to 0.974	0.545 to 0.880	0.582 to 0.906	0.690 to 0.961	0.500 to 0.848	0.664 to 0.949	0.706 to 0.968	0.523 to 0.865	0.622 to 0.927	0.616 to 0.924	0.614 to 0.923	0.614 to 0.923	0.693 to 0.963
Z-test	72.8	9.507	4.491	4.538	11.189	2.590	6.220	15.083	10.074	5.197	5.120	5.120	3.066

Table 6. Comparison of physician and Swin performance on 30 test images. Metrics are averaged across classes.

dataset used in this study included images from both public sources (e.g., DermNet) and Thai patients, we did not conduct a formal ablation analysis to isolate the impact of dataset origin on model performance. Informal observations during model development indicated that networks trained solely on lighter-skinned datasets (e.g., DermNet) yielded reduced accuracy on images from Thai patients, particularly in conditions such as eczema, where erythema presents less prominently. This underscores the need for training datasets that reflect a diversity of skin tones and lesion morphologies to ensure generalizability. In contrast, non-dermatologists, whose clinical training primarily involves history taking and physical examination, were limited in this study to interpreting photographic images without access to contextual patient histories or additional diagnostic tools. This reliance on static images highlights the advantage of data-driven deep learning models in standardizing and enhancing diagnostic accuracy. Similar insights were reported in a study by Liu et al., which demonstrated that a deep learning system achieved diagnostic accuracy comparable to dermatologists and surpassed that of primary care physicians and nurse practitioners in classifying skin conditions[46]. Additionally, a study by Venkatesh et al., found that deep learning models exhibited higher diagnostic accuracy than non-dermatologists in dermatology, further supporting the collaborative potential of AI in clinical workflows[47].

While internists outperformed interns, reflecting the value of clinical experience, their diagnostic accuracy remained below that of Swin, particularly for conditions with complex presentations like eczema. This discrepancy highlights the need for structured dermatology training to ensure high diagnostic accuracy.

The prevalence of dermatophytosis (20–25%) and eczema (up to 20%) compared to psoriasis (1–10%) may influence diagnostic familiarity among non-dermatologists in real-world settings[1,48]. Despite this, eczema remains the most challenging condition for non-dermatologists due to its varied clinical presentations.

From the confusion matrix, interns show considerable difficulty in diagnosing eczema, with 22% of eczema cases being misclassified as dermatophytosis and 10% as psoriasis. While interns perform relatively well in identifying psoriasis (85%) and dermatophytosis (81%), the clinical variability of eczema introduces significant errors. In contrast, internists demonstrate improved diagnostic accuracy, with 71% correct predictions for eczema, though 20% of eczema cases are still misclassified as psoriasis. The challenges with eczema persist even for experienced internists, highlighting its complexity in clinical practice. Figure 11 shows the confusion matrix for novices and experts in diagnosing three skin diseases)

The ability of Swin to consistently outperform non-dermatologists across all conditions, supported by Grad-CAM visualizations that enhance interpretability, underscores its potential as a transformative tool in dermatology. These visualizations enhance clinician trust and facilitate AI integration into clinical workflows. Furthermore, the challenges highlighted in this study, including reliance on photographic data and limited contextual information, advocate for integrating deep learning models into dermatological training to complement clinical expertise and improve patient outcomes. To translate these findings into practical deployment, the Swin could serve as a triage or decision-support tool in telemedicine,assisting primary care providers in identifying high-risk cases or confirming suspected diagnoses. For real-world deployment, model explainability, clinician oversight, and medicolegal frameworks are essential to ensure safety and accountability. Clear governance protocols and human-in-the-loop safeguards should be established to address diagnostic liability and maintain clinician trust. These elements will be critical for transitioning AI systems like the Swin from research to clinical implementation. These steps are essential to move from proof-of-concept to reliable and safe adoption in everyday practice. In real-world workflows, the Swin's interpretability via Grad-CAM could be integrated as a visual overlay during telemedicine consultations or electronic health record systems, allowing physicians to cross-check AI focus with clinical features. Such interpretability not only enhances clinician trust but also provides educational value for trainees. To ensure patient safety, a human-in-the-loop framework is envisioned, where ambiguous or low-confidence cases trigger clinician review, and user feedback on AI-assisted decisions is collected to iteratively refine the model's deployment. The model's robustness and interpretability make it an appealing candidate for deployment in general practice and telemedicine workflows. However, the current study did not evaluate performance under real-world telemedicine conditions, such as uncontrolled lighting, motion blur, or low-resolution images from patient-owned devices. These variables may impact classification reliability and should be investigated in prospective validation. Additionally, the evaluation of human performance was conducted as a pilot study using 30 clinical images assessed by 30 non-dermatologist participants. While this design allowed for a direct comparison with the deep learning models, its limited sample size may restrict statistical power and reduce the generalizability of findings. The human–AI comparison in this study was intentionally designed as a pilot, using 30 clinical images and 30 non-dermatologist participants. While this design allowed for a controlled, qualitative benchmark, its limited scale reduces statistical power and restricts the generalizability of conclusions. Therefore, these findings should be considered exploratory and hypothesis-generating rather than definitive. To address this limitation, future research will incorporate larger sample sizes, a broader range of evaluator expertise, and more heterogeneous image sets to strengthen the reliability and applicability of comparative analyses.

By incorporating findings from comparative studies, this discussion underscores the potential of Swin and similar deep learning models to enhance diagnostic workflows in dermatology while addressing existing challenges in clinical and AI integration

Strengths and limitations

This study offers several notable strengths. First, it utilized a diverse dataset drawn from a public source, DermNet[9] and a local Thai clinical dataset. This diversity enhances the model's generalizability and ensures representation across a wide range of dermatologic conditions. Second, the inclusion of Asian patients, who generally have darker skin types than Caucasians[1], increases the relevance of our findings for populations that are often underrepresented in dermatologic AI studies. Third, all clinical images were captured using various smartphone models under realistic conditions, reflecting the quality and variability typical of teledermatology environments. Lastly, model interpretability was addressed using Grad-CAM, which demonstrated that the model consistently attended to clinically meaningful regions, providing transparency that may enhance clinician trust and educational utility. Nevertheless, this study has limitations. Differences in skin pigmentation can alter the visual characteristics of dermatologic lesions[49], and both deep learning models and physicians have shown reduced diagnostic accuracy in darker-skinned populations[50]. Additionally, to comply with the Thai Personal Data Protection Act, we excluded lesions from the face, neck, and groin. While this approach was necessary for ethical and legal compliance, it restricts the generalizability of our model to clinically important areas, such as inverse psoriasis, seborrheic dermatitis, and intertriginous eczema[51]. These anatomical sites often pose diagnostic challenges due to overlapping morphologies, and their exclusion represents a meaningful limitation of this work. Future studies should aim to incorporate such regions through carefully designed, privacy-compliant protocols to enhance applicability in real-world practice. The human-AI comparison was conducted as a pilot study with 30 clinical images reviewed by 30 non-dermatologists. While this design provided an initial benchmark, the limited sample size reduces statistical power and generalizability. Additionally, the study did not evaluate medicolegal risks associated with AI misdiagnosis or the operational implications of deploying such models in clinical practice. This study did not evaluate performance under real-world telemedicine conditions, such as uncontrolled lighting, motion blur, or low-resolution images from patient-owned devices. Device variability, including differences in smartphone models, camera quality, and ambient lighting, can substantially affect image appearance and diagnostic reliability. These factors limit direct applicability of our results to telemedicine environments, underscoring the need for future validation under uncontrolled, real-world imaging conditions.

Future research should include larger, more diverse clinician cohorts and broader anatomical coverage, validated under real-world telemedicine conditions. Moreover, efforts should be made to stratify performance across skin phototypes and establish clinical oversight frameworks for safe AI deployment, including mechanisms to flag uncertain or ambiguous cases. These steps are essential to ensure both the scalability and safety of AI-assisted dermatologic diagnosis in routine care.

Conclusions

This study developed a deep learning framework leveraging CNN and Transformer architectures to classify dermatophytosis, psoriasis, and eczema. Swin outperformed all models, demonstrating the highest accuracy and F1 scores, minimal misclassification, and interpretable predictions via Grad-CAM, enhancing its clinical applicability.

Swin also surpassed non-dermatologists in diagnostic performance, particularly for challenging conditions like eczema, highlighting its potential as a diagnostic aid in primary care and telemedicine. The model's robustness across diverse datasets, including Thai skin phototypes, underscores its suitability for varied populations, though its exclusion of facial, neck, and groin lesions limits generalizability.

These findings support integrating AI tools like the Swin Transformer into clinical practice to enhance diagnostic accuracy and educate non-specialists. Further large-scale validation across diverse populations is warranted.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 10 May 2025; Accepted: 18 November 2025
Published online: 04 December 2025

References

1. Kimball, A. B. Skin differences, needs, and disorders across global populations. In *Journal of Investigative Dermatology Symposium Proceedings* **13**, 2–5. https://doi.org/10.1038/jidsymp.2008.5 (2008).
2. Egeberg, A., Griffiths, C., Williams, H., Andersen, Y. & Thyssen, J. Clinical characteristics, symptoms and burden of psoriasis and atopic dermatitis in adults. *Br. J. Dermatol.* **183**, 128–138 (2020).
3. Chong, M. & Fonacier, L. Treatment of eczema: Corticosteroids and beyond. *Clin. Rev. Allergy Immunol.* **51**, 249–262 (2016).
4. Gnat, S., Łagowski, D. & Nowakiewicz, A. Major challenges and perspectives in the diagnostics and treatment of dermatophyte infections. *J. Appl. Microbiol.* **129**, 212–232 (2020).
5. Gottlieb, A. B. & Dann, F. Comorbidities in patients with psoriasis. *Am. J. Med.* **122**, 1150-e1 (2009).
6. Basarab, T., Munn, S. & Jones, R. R. Diagnostic accuracy and appropriateness of general practitioner referrals to a dermatology out-patient clinic. *Br. J. Dermatol.* **135**, 70–73 (1996).
7. Liu, Z., Wang, X., Ma, Y., Lin, Y. & Wang, G. Artificial intelligence in psoriasis: Where we are and where we are going. *Exp. Dermatol.* **32**, 1884–1899 (2023).
8. Zhang, J. et al. Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review. *Diagnostics* **13**, 3506 (2023).
9. Hammad, M., Pławiak, P., ElAffendi, M., El-Latif, A. A. A. & Latif, A. A. A. Enhanced deep learning approach for accurate eczema and psoriasis skin detection. *Sensors* **23**, 7295 (2023).
10. Krakowski, I. et al. Human-ai interaction in skin cancer diagnosis: a systematic review and meta-analysis. *NPJ Digit. Med.* **7**, 78 (2024).
11. Hameed, N., Shabut, A. M., Ghosh, M. K. & Hossain, M. A. Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Syst. Appl.* **141**, 112961 (2020).
12. Huang, K. et al. The classification of six common skin diseases based on xiangya-derm: Development of a chinese database for artificial intelligence. *J. Med. Internet Res.* **23**, e26025 (2021).
13. Eskandari, A. & Sharbatdar, M. Efficient diagnosis of psoriasis and lichen planus cutaneous diseases using deep learning approach. *Sci. Rep.* **14**, 9715 (2024).
14. Wu, H. et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Ann. Transl. Med.* **8**, 581 (2020).

15. Hammad, M., Plawiak, P., ElAffendi, M., El-Latif, A. A. A. & Latif, A. A. A. Enhanced deep learning approach for accurate eczema and psoriasis skin detection. *Sensors* https://doi.org/10.3390/s23167295 (2023).
16. Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285 (2016).
17. Taylor, L. & Nitschke, G. Improving deep learning using generic data augmentation. arXiv preprint arXiv:1708.06020 (2017).
18. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114 (PMLR, 2019).
19. Krizhevsky, A., Sutskever, I. & Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.***25** (2012).
20. Krizhevsky, A. One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997 (2014).
21. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708 (2017).
22. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, 10096–10106 (PMLR, 2021).
23. Szegedy, C. et al. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
24. Koonce, B. & Koonce, B.E. *Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization* (Springer, 2021).
25. Iandola, F.N. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016).
26. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
27. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
28. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
29. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021).
30. Wu, H. et al. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22–31 (2021).
31. Ding, M. et al. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, 74–92 (Springer, 2022).
32. Tu, Z. et al. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*, 459–479 (Springer, 2022).
33. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J. & Molchanov, P. Global context vision transformers. In *International Conference on Machine Learning*, 12633–12646 (PMLR, 2023).
34. Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O. & Ranjan, A. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5785–5795 (2023).
35. Yun, S. & Ro, Y. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5756–5767 (2024).
36. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (IEEE, 2009).
37. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inform. Process. Manage.* **45**, 427–437 (2009).
38. Chinchor, N.A. Muc-4 evaluation metrics. In *Message Understanding Conference* (1992).
39. Powers, D.M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020).
40. Mohan, J. et al. Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai. *Comput. Biol. Med.* **190**, 110007 (2025).
41. Adebiyi, A. et al. Transformers in skin lesion classification and diagnosis: A systematic review. *medRxiv* **09**, 2024 (2024).
42. Ahmad, J., Farooq, M.U., Zafeer, F., Shahid, N. et al. Classification of 24 skin conditions using swin transformer: Leveraging dermnet & healthy skin dataset. In *International Conference on Energy, Power, Environment, Control and Computing (ICEPECC 2025)*, vol. 2025, 511–519 (IET, 2025).
43. Saber, A., Elbedwehy, S., Awad, W. A. & Hassan, E. An optimized ensemble model based on meta-heuristic algorithms for effective detection and classification of breast tumors. *Neural Comput. Appl.* **37**, 4881–4894 (2025).
44. Elbedwehy, S., Hassan, E., Saber, A. & Elmonier, R. Integrating neural networks with advanced optimization techniques for accurate kidney disease diagnosis. *Sci. Rep.* **14**, 21740 (2024).
45. Alnowaiser, K., Saber, A., Hassan, E. & Awad, W. A. An optimized model based on adaptive convolutional neural network and grey wolf algorithm for breast cancer diagnosis. *PLoS One* **19**, e0304868 (2024).
46. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908. https://doi.org/10.1038/s41591-020-0842-3 (2020).
47. Venkatesh, K. P., Raza, M. M., Nickel, G., Wang, S. & Kvedar, J. C. Deep learning models across the range of skin disease. *npj Dig. Med.* **7**, 32 (2024).
48. Chanyachailert, P., Leeyaphan, C. & Bunyaratavej, S. Cutaneous fungal infections caused by dermatophytes and non-dermatophytes: An updated comprehensive review of epidemiology, clinical presentations, and diagnostic testing. *J. Fungi* **9**, 669 (2023).
49. Myers, J. Challenges of identifying eczema in darkly pigmented skin. *Nurs. Child. Young People* **27**, 24–28. https://doi.org/10.7748/ncyp.27.6.24.e571 (2015).
50. Groh, M. et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nat. Med.* **30**, 573–583. https://doi.org/10.1038/s41591-023-02728-3 (2024).
51. Micali, G. et al. Inverse psoriasis: From diagnosis to current treatment options. *Clin. Cosmet. Investig. Dermatol.* **12**, 953–959. https://doi.org/10.2147/CCID.S189000 (2019).

Acknowledgements

The authors gratefully acknowledge the patients who participated in this study; the clinicians who participated in the tests; Dr. Prapat Suriyaphol for his advice about the Personal Data Protection Act in Thailand, Dr. Irin Chaikangwan for her advice about deep learning, Ms. Nutwara Meannui for her assistance with data collection, and Dr. Chulaluk Komoltri for her advice with statistical analysis. This research project was supported by Faculty of Medicine Siriraj Hospital, Mahidol University, Grant Number (IO) R016633016.

Declaration of generative AI and AI-assisted technologies in writing process

During the preparation of this work, the authors used ChatGPT and Grammarly to improve readability and grammar. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Author contributions

N.Y.: Conceptualization, acquisition of data, analysis or interpretation of data, Writing—Original Draft, Writing—Review & Editing, administrative, technical, or material support, supervision. C.W.: Conceptualization, acquisition of data, analysis or interpretation of data, drafting of the manuscript, administrative, technical, or material support, supervision. T.T.: Conceptualization, Acquisition of data, Data analysis and interpretation, Writing—Original Draft, Writing—Review & Editing ,administrative, technical, or material support, Supervision. L.C.: Conceptualization, acquisition of data, analysis or interpretation of data, Writing—Original Draft, administrative, technical, or material support. N.S.: Conceptualization, acquisition of data, analysis or interpretation of data, Writing—Original Draft, administrative, technical, or material support. C.C.: Conceptualization, acquisition of data, analysis or interpretation of data, Writing—Original Draft, administrative, technical, or material support. S.B.: Conceptualization, acquisition of data, analysis or interpretation of data. T.N.: Conceptualization, acquisition of data, analysis or interpretation of data. T.P.: Conceptualization, acquisition of data, analysis or interpretation of data. T.H.: Acquisition of data, analysis or interpretation of data. P.W.: Acquisition of data, analysis or interpretation of data. P.K.: Acquisition of data, analysis or interpretation of data. S.A.: Conceptualization, Acquisition of data, Data Analysis and Interpretation, Writing—Original Draft, Writing—Review & Editing, Supervision P.C.: Conceptualization, acquisition of data, analysis or interpretation of data, Writing—Original Draft, Writing—Review & Editing, administrative, supervision. administrative, technical, or material support.

Funding

This research project was supported by Faculty of Medicine Siriraj Hospital, Mahidol University, Grant Number (IO) R016633016.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-29562-6.

Correspondence and requests for materials should be addressed to P.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.