scientific reports

OPEN

Artificial intelligence-assisted prediction of *Demodex* mite density in facial erythema

Jemin Kim[1,7], Yun Na Lee[2,7], Jihee Boo[1], Inrok Oh[3], Changyoon Lee[4], Joo Hee Lee[5], Ye Seul Choi[1], Hyun Kim[1], Jung Im Na[6], Jihee Kim[1]✉ & Chang Ook Park[5]✉

Current detection methods of *Demodex* mite density in facial erythema are semi-invasive or operator-dependent. We developed and evaluated a deep learning model (DemodexNet) for predicting *Demodex* mite density and assessed its impact on the diagnostic performance of dermatologists. This study included 1,124 patients with facial erythema who underwent *Demodex* mite density measurement at two referral hospitals between January 2016 and August 2023. DemodexNet achieved area under the receiver operating characteristic curve values of 0.823–0.865 in internal testing, with lower values observed in the external testing set. AI-assisted evaluation was associated with an increase in diagnostic accuracy among dermatologists from 63.7% to 70.6% ($P < .001$). Less experienced dermatologists and those with higher trust in AI showed greater performance gains. The model recognized central facial regions and individual lesions characteristic of demodicosis. DemodexNet demonstrates promising performance in predicting *Demodex* mite density and significantly improves dermatologists' diagnostic accuracy. As this proof-of-concept study was limited to Korean patients with Fitzpatrick skin types III-IV, validation in diverse populations is required before broader clinical application.

Keywords Facial erythema, Rosacea, *Demodex* mites, Artificial intelligence, Deep learning

Facial erythema, often referred to as "red face," is a readily identifiable clinical manifestation in dermatology[1]. However, its presentation can result from diverse dermatological diseases and other medical conditions[2]. Rosacea is one of the most common chronic inflammatory conditions that present with frequent flushing and facial erythema. This condition can greatly affect the quality of life of patients and is associated with an increased risk of cardiovascular, gastrointestinal, mental, and neurological problems[3]. Although rosacea is the most emblematic disease associated with "red face," the differential diagnosis encompasses a broad spectrum of conditions, including contact dermatitis, atopic dermatitis, seborrheic dermatitis, acne vulgaris, lupus erythematosus, and dermatomyositis[4,5].

Various genetic, environmental, and microbial determinants have been implicated in the etiology of facial erythema. Among these, *Demodex* mites, which reside within the pilosebaceous units of human facial skin as commensals, are significant contributors[6]. *Demodex* mites are commonly found in the skin of healthy adults, with a prevalence rate of 100% and a density of ≤ 5 mites/cm^2[7]. However, overproliferation of these mites can lead to pathogenic processes referred to as demodicosis, which result in various symptoms, including facial redness, irritation, itching, and inflammation[8,9]. The diagnosis of demodicosis is based on the evaluation of the number of *Demodex* mites present on the skin surface. This can involve a standardized skin surface biopsy or direct microscopic examination of fresh secretions from sebaceous glands (DME)[10,11]. However, the utility of these methodologies in routine clinical practice is limited owing to their painful, semi-invasive nature and the significant impact of operator proficiency on the results[10,12].

This study aimed to develop a deep learning model called DemodexNet, which can predict the density of *Demodex* mites by analyzing clinical data and photographs of patients with facial erythema. Furthermore,

[1]Department of Dermatology, Yongin Severance Hospital Yonsei University College of Medicine, Gyeonggi-do 16995, Yongin-si, Korea. [2]Department of Transdisciplinary Medicine , Seoul National University Hospital , Seoul 03080, Korea. [3]LG Chem Ltd , Seoul, Korea. [4]Department of Medicine , Yonsei University College of Medicine , Seoul 03722, Korea. [5]Department of Dermatology and Cutaneous Biology Research Institute , Severance Hospital Yonsei University College of Medicine , Seoul 03722, Korea. [6]Department of Dermatology, Seoul National University Bundang Hospital, Gyeonggi-do 13620, Seongnam-si, Korea. [7]Jemin Kim and Yun Na Lee authors contributed equally to this work. ✉email: mygirljihee@yuhs.ac; COPARK@yuhs.ac

the study assessed the effectiveness of the model in enhancing the ability of dermatologists to identify overproliferation of *Demodex* mites.

Methods

Study design and participant selection

This diagnostic study was approved by the Institutional Review Board of Yonsei University Severance Hospital and Yongin Severance Hospital (approval numbers 4–2023-1008 and 2023-0382-001, respectively). The study adheres to the Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology[34] and the Standards for Reporting of Diagnostic Accuracy Studies reporting guidelines. The requirement for informed consent was waived because retrospective and deidentified data were used.

This study included all patients diagnosed with facial erythema between January 2016 and August 2023 who underwent *Demodex* mite density measurement at two referral hospitals in South Korea: Severance Hospital and Yongin Severance Hospital. The *Demodex* mite density was quantified using the DME method, as previously described[9,35]. A density of > 5 mites/cm^2 was classified as high (positive) *Demodex* infestation[10]. Patients were excluded if frontal facial photographs were not obtained during their clinic visit on the day of the *Demodex* examination.

Data preparation and preprocessing

The study included digital images of the face, which underwent an automated face detection and deidentification process to protect personally identifiable information. Using the Mediapipe library, facial landmark coordinates were extracted, and polygonal masks were drawn over the eyes and mouth to protect anonymity. To address class imbalance and prevent overfitting, data augmentation was applied, including geometric transformations and color-based enhancements. Full procedures are detailed in Supplementary Methods S1. Several approaches for handling class imbalance were compared, with data augmentation chosen as the preferred method (see Supplementary Table S5).

Additionally, clinical data were collected for each patient, including age, sex, clinical symptoms (itching, burning or stinging, edema, dryness, and flushing), serum allergy marker levels (eosinophil cationic protein, total immunoglobulin E (IgE), and eosinophil counts), patch test results, and the presence of extra-facial skin lesions[32]. Missing values in serum allergy markers (19.0% of the dataset) were imputed using the MissForest algorithm with optimized hyperparameters as detailed in Supplementary Methods S4.

Model development

To develop an AI model that learns the distribution of facial demodicosis and individual localized lesions situated around complex anatomical landmarks while integrating clinical information, we implemented a two-fold strategy in our model development process. First, a stacking ensemble (SE) model was developed, layering networks focusing on the comprehensive facial image and localized patches indicative of *Demodex* infestation (Fig. 1a)[20,21]. Second, we applied the Globally-aware Multiple Instance Classifier (GMIC), a weakly supervised model designed for end-to-end training to independently identify patches associated with *Demodex* infestation from the full image (Fig. 1b)[23,36]. While the architectures of these models are independent and differ, both include global and local modules for capturing the nuanced features of *Demodex* infestation. Moreover, each model incorporates 12 clinical variables related to the image data in a distinct module, leading to a combined prediction model that merges insights from both image and clinical data (Supplementary Figure S1). The data collected from Severance Hospital constituted the primary dataset, with an allocation of 80% for model training, 10% for validation, and 10% for internal testing. The entire dataset from Yongin Severance Hospital (100%) was designated as the external testing set (Table 1). Additionally, we employed 10-fold stratified cross-validation to maintain model performance robustness.

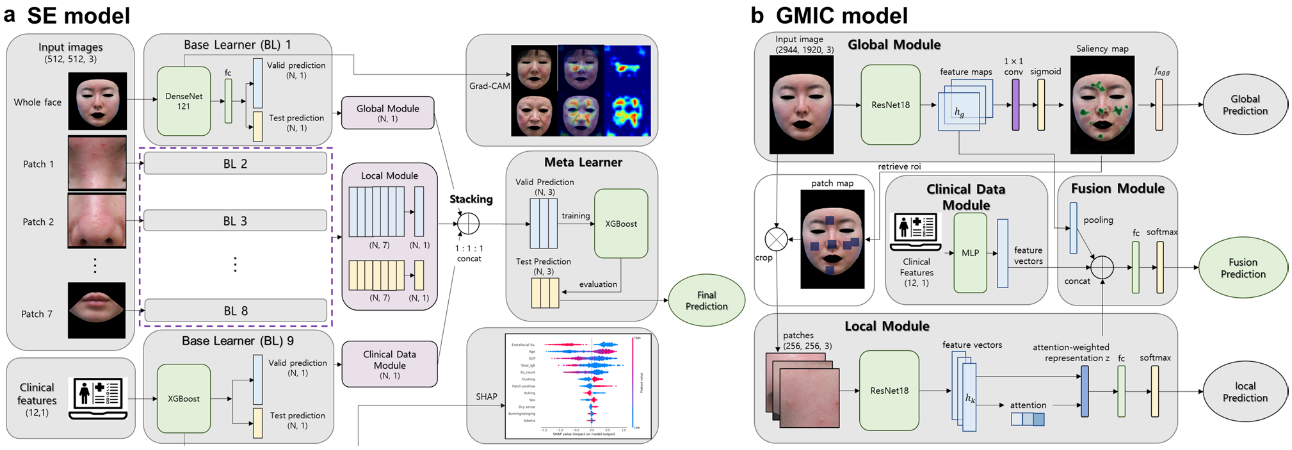

Fig. 1. Overview of the DemodexNet model architecture. (a) Stacking ensemble (SE) model, (b) Globally-aware Multiple Instance Classifier (GMIC) model.

Characteristics	Dataset	
	Main	External
Data collection period	2016. 1–2022. 12	2020. 3–2023. 8
Location (hospital)	Department of dermatology, Severance Hospital	Department of dermatology, Yongin Severance Hospital
Dataset allocation	Training (80%)	External testing (100%)
	Validation (10%)	
	(Internal) Testing (10%)	
Camera type	Digital camera (Canon EOS RP 24–105 mm; 26.2 megapixels)	Digital camera (Canon EOS 800D, 24.2 megapixels)
Lighting condition	Standardized indoor clinical lighting with a uniform blue background and white overhead illumination	
Patient demographics		
Unique individuals, n	1024	100
Female sex	697 (68.1)	70 (70.0)
Age at diagnosis	32.5 (24.0–47.0)	37.5 (28.0–53.0)
High *Demodex* density	255 (24.9)	45 (45.0)
Associated symptoms[a]		
Flushing	434 (42.4)	49 (49.0)
Itching	828 (80.9)	78 (78.0)
Burning/stinging	223 (21.8)	30 (30.0)
Edema	127 (12.4)	21 (21.0)
Dry sense	154 (15.0)	19 (19.0)
Positive patch test	316 (30.9)	41 (41.0)
Extrafacial skin involvement	490 (47.9)	26 (26.0)
Serum allergy marker		
ECP (μg/L)	26.2 (17.0–41.0)	22.7 (16.2–31.7)
Eosinophil count (cells/μL)	160.0 (80.0–310.0.0.0)	111.2 (67.5–205.0)
Total IgE (IU/mL)	111.2 (35.0–461.9.0.9)	66.7 (27.7–205.0)

Table 1. Summary of the main and external datasets. ECP, eosinophil cationic protein; IgE, immunoglobulin E. Data are presented as n (%) or median (interquartile range)[a]Patients might be listed in >1 category

SE architecture

The SE framework comprises global, local, and clinical modules, which take as input a whole facial image $\mathrm{G} \in \mathbb{R}^{H\times W\times 3}$, seven localized patches $\mathrm{P}_k \in \mathbb{R}^{H\times W\times 3}$ with $k = 1, \ldots, 7$, and clinical data $\mathrm{C} \in \mathbb{R}^{12}$. We extracted patches, including the forehead, nose, cheeks, and chin, based on facial landmark coordinates obtained during the deidentification process. Each module employed an independent feature extractor, referred to as a base learner: f_g and f_l (DenseNet121 for image modules) and f_c (XGBoost for clinical data). The global module processes the whole face image G using a DenseNet121 and outputs a probability as

$$y_g = f_g\left(G\right)$$

Similarly, the local module processes each patch P_k to output $y_{l,k} = f_l\left(P_k\right)$, and the outputs from the local module are then summed to obtain the integrated local score:

$$y_l = \sum\nolimits_{k=1}^{7} y_{l,k}$$

The clinical module produces $y_c = f_c\left(C\right)$. These outputs are concatenated into a feature vector $\mathrm{z} = [y_g,\ y_l,\ y_c]$, which is subsequently passed to a meta learner f_m to obtain the final prediction:

$$\widehat{y}_{se} = f_m\left(z\right)$$

Each base learner is trained using the binary cross-entropy loss:

$$L_b = BCE = -\frac{1}{N}\sum\nolimits_{i=1}^{N}\left[y_i \log\widehat{y_i} + (1-y_i)\log(1-\widehat{y_i})\right]$$

Additional implementation details are provided in Supplementary Methods S2.

GMIC architecture

The GMIC also consists of global, local, and clinical modules. It takes as input the original high-resolution image $\mathrm{x} \in \mathbb{R}^{H\times W\times 3}$ along with clinical data C. A global network f_g first extracts a feature map as follows:

$$h_g = f_g\left(x\right)$$

Then, a 1×1 convolution followed by a sigmoid activation produces a saliency map A, which highlights regions potentially relevant to *Demodex* infestation. Based on the saliency map, K regions of interest (ROIs), representing the most informative patches, are selected from the input x according to the following procedure:

$$\left\{ \tilde{x}_k \right\} = retrieve\_roi\left(A\right)$$

We employed a greedy algorithm to retrieve K patches, denoted as $\tilde{x}_k \in \mathbb{R}^{h_c, w_c}$, where we set $w_c = h_c = 256$. We heuristically set K =6 (see Supplementary Table S6), and each patch was then processed by a local network f_l to obtain feature vectors:

$$\tilde{h}_k = f_l\left(\tilde{x}_k\right)$$

Attention weights $\alpha_{\,k}$ are computed using a gated mechanism, as the selected ROI patches do not contribute equally to the final prediction. This mechanism enables the model to assign higher weights to more informative patches, thereby enabling effective aggregation of local features:

$$\alpha_{\,k} = \frac{\exp\left\{ w^T \left(\tanh\left(V\tilde{h}_k^\top\right) \odot sigm\left(U\tilde{h}_k^\top\right)\right)\right\}}{\sum_{j=1}^{K} \exp\left\{ w^T \left(\tanh\left(V\tilde{h}_j^\top\right) \odot sigm\left(U\tilde{h}_j^\top\right)\right)\right\}}$$

with learnable parameters $\mathbf{w} \in \mathbb{R}^L$, $\mathbf{V} \in \mathbb{R}^{L \times M}$, and $\mathbf{U} \in \mathbb{R}^{L \times M}$ with $\mathrm{L} = 512$ and $\mathrm{M} = 128$. The weight sum via attention-weighted aggregation,

$$\mathrm{z} = \sum_{k=1}^{K} \alpha_{\,k} \tilde{h}_k$$

is used to represent the locally aggregated feature.

Clinical data are separately encoded by a multi-layer perceptron to produce the feature vector $h_c = f_c\left(C\right)$. To integrate the global and local feature, we applied global max pooling on h_g and concatenated it with z and h_c. The fused representation is subsequently passed through a fully connected layer with a softmax activation to yield the final prediction:

$$\hat{y}_{gmic} = \hat{y}_{fusion} = soft\max(w_f\left[GMP\left(h_g\right),\ z\right]^\top$$

where $\boldsymbol{w}_f$ values are learnable parameters. The GMIC is trained with the following loss function:

$$L_{gmic} = \sum BCE_{local} + BCE_{global} + BCE_{fusion} + \beta\, L_{reg}\left(A\right)$$

where β is a weighting coefficient. To encourage the saliency map to focus only on highly informative regions, we applied the L1 regularization on A:

$$L_{reg}\left(A\right) = \sum_{(i,j)} |A_{i,j}|$$

Further implementation details are provided in Supplementary Methods S3.

Human evaluators and decision study

Twenty-one participants, comprising 10 dermatology residents and 11 board-certified dermatologists, were recruited to evaluate the ability of human evaluators to classify *Demodex* infestation cases using facial images and clinical data. The study also aimed to assess potential performance improvement with the assistance of DemodexNet. An anonymous online questionnaire was administered in two parts over a two-week interval using Google Survey (Supplementary Figure S2).

We used all 100 cases from the internal testing dataset in Part I. Participants were presented with original-resolution photographs and associated clinical data (age, sex, clinical symptoms, and the presence of extra-facial skin lesions). They were asked to classify each case as positive or negative for *Demodex* infestation. In Part II, participants were given DemodexNet's confidence scores (ranging from 0 to 1) for each case, with scores ≥ 0.5 indicating a model prediction of *Demodex* positivity. Participant performance was assessed by comparing their predictions with the gold standard label. Case sequences were shuffled between parts to ensure unbiased responses. Reference diagnoses and participant scores were not disclosed until the conclusion of the study.

Evaluation of algorithm performance and statistics

Model performance was evaluated using top-1 accuracy, sensitivity, specificity, and the F1 score. Receiver operating characteristic (ROC) curves were plotted using sensitivity and specificity for each threshold, and areas under the curve (AUCs) were calculated. Additionally, 95% confidence intervals (CIs) were obtained via nonparametric bootstrap of predictions with replacement ($N = 1000$), using the same sample size as the internal test set and external set for each resample[37].

For a visual explanation of *Demodex* mite distribution predictions, we implemented gradient-weighted class activation mapping (Grad-CAM)[38] on each base learner for the SE model. For the GMIC model, we visualized saliency maps based on the algorithm's inherent attention scores. To identify significant variables in predicting algorithm outcomes for the clinical-data-based model, we employed SHapley Additive exPlanations (SHAP) to visualize feature importance rankings[39]. Additionally, we conducted multivariate logistic regression analyses to identify predictive factors associated with high *Demodex* mite density.

We determined the accuracy, sensitivity, and specificity of human participants for each part, comparing their performances before and after algorithm assistance using the McNemar test. This test was specifically chosen for paired nominal data, where each evaluator assessed the same 100 images twice. We constructed 2 × 2 contingency tables for each evaluator comparing their diagnostic decisions (correct/incorrect) between Part I (image only) and Part II (image with AI assistance) to evaluate whether AI assistance significantly improved diagnostic performance.

Fleiss' kappa (κ) values were calculated to evaluate agreement among participants' responses, and we generated a heatmap using hierarchical agglomerative clustering to visualize interparticipant agreement rates[24,40]. Statistical analyses were performed using Python version 3.9.7 and R version 4.1.3. Statistical significance was set at a two-tailed *P*-value < 0.05.

Results

Study dataset

The study included 1,024 and 100 patients in the main and external datasets, respectively. The main dataset was collected at Severance Hospital from January 2016 to December 2022, while the external dataset was obtained at Yongin Severance Hospital from March 2020 to August 2023. Both sites used standardized indoor clinical lighting with a consistent blue background and white overhead illumination, along with similar digital cameras. The datasets showed comparable gender distributions (68.1% and 70.0% female, respectively) but differed in median age at diagnosis (32.5 vs. 37.5 years) and notably in *Demodex* positivity rates (24.9% vs. 45.0%). Detailed participant characteristics are summarized in Table 1, and baseline comparisons of clinical features by *Demodex* mite density are shown in Supplementary Table S1.

The study included 1,024 and 100 patients in the main and external datasets, respectively. Both datasets predominantly comprised female participants (68.1% and 70.0%, respectively), with median ages at diagnosis of 32.5 and 37.5 years, respectively.

Participant characteristics are summarized in Table 1, and baseline comparisons of clinical features by *Demodex* mite density are shown in Supplementary Table S1.

Model performance

Table 2 summarizes the performance metrics of the DemodexNet models—sensitivity, specificity, F1 score, ROC–AUC, and accuracy—reported from the fold with the smallest absolute validation–test ROC–AUC gap. For the SE model, the image-based model on the internal testing set achieved an ROC–AUC of 0.825 (95% CI: 0.734–0.903), with relatively high specificity (0.980 [95% CI: 0.937–1.000]) and low sensitivity (0.260 [95% CI: 0.143–0.380]). The clinical data-based model achieved an ROC–AUC of 0.842 (95% CI: 0.751–0.915), while the combination of these two models yielded an ROC–AUC of 0.823 (95% CI: 0.728–0.896). For the GMIC model, the image-based model achieved an ROC-AUC of 0.833 (95% CI: 0.753–0.908) on the internal testing set, with balanced specificity (0.760 [95% CI: 0.633–0.872]) and sensitivity (0.640 [95% CI: 0.500–0.767]). The clinical data-based model achieved an ROC–AUC of 0.790 (95% CI: 0.680–0.873), while the combined model yielded an improved ROC–AUC of 0.865 (95% CI: 0.785–0.934). Both models exhibited similar trends in the external test dataset, although with lower ROC–AUC and accuracy values than the internal test set results (Supplementary Figure S3).

Augmented decision-making with artificial intelligence

We invited 21 dermatologists to participate in a two-step reader study to validate the decision support provided by DemodexNet. Without AI assistance, participants demonstrated an accuracy of 0.637 (95% CI: 0.615–0.656), representing an absolute improvement of 6.9% (95% CI: 4.1–9.7%, *P*<.001). The effect was most pronounced for sensitivity, with an absolute increase of 13.6% (95% CI: −16.6–16.6%, *P*<.001), while specificity remained relatively stable with a 0.2% change (95% CI: −2.6–2.6%, *P*=.95).

When analyzing human raters divided into three subgroups based on clinical experience, the magnitude of improvement varied notably. The low-experience group (> 2 years) showed the most significant benefit, with an 11.6% absolute accuracy improvement, while the high-experience group (> 8 years) demonstrated a 5.8% improvement; both achieved statistical significance (*P*<.001 and *P*=.01, respectively). The low-experience group initially underperformed compared with the AI model in Part I. However, in Part II, this group showed more pronounced improvements in accuracy, sensitivity, and specificity than more experienced groups (Fig. 2a and Supplementary Table S2). Comparing responses between Parts I and II to assess changes in rater response based on AI predictions, we observed that less experienced raters were more likely to modify their responses to correct answers when assisted by AI. Notably, this increased correction rate was not accompanied by a higher tendency to follow incorrect AI predictions (Fig. 2b).

Following Part II, we conducted a 5-point questionnaire-based assessment of DemodexNet among human evaluators[13]. On the basis of the survey results, we divided participants into two subgroups: those with a positive impression of DemodexNet (Trust group) and those without (Untrust group). We then performed a subgroup analysis (Supplementary Figure S4). Results showed that the Trust group demonstrated significantly higher positive benefits when modifying their responses according to AI predictions. Notably, the two groups had no significant difference in responses to incorrect AI predictions (Fig. 2c).

	Classification performance (95% CI)[a]					
Model (class)	**Sensitivity**	**Specificity**	**F1-Score**	**ROC-AUC**	**Accuracy**	***P*-value[b]**
SE model						
Internal testing set						
Image-based model	0.260 (0.143–0.380)	0.980 (0.937–1.000)	0.406 (0.255–0.543)	0.825 (0.734–0.903)	0.620 (0.520–0.720)	0.97
Clinical-data-based model	0.840 (0.731–0.935)	0.780 (0.660–0.894)	0.816 (0.725–0.889)	0.842 (0.751–0.915)	0.810 (0.720–0.880)	0.74
Combined model	0.300 (0.226–0.371)	0.960 (0.927–0.987)	0.448 (0.362–0.529)	0.823 (0.728–0.896)	0.630 (0.573–0.683)	Ref
External testing set						
Image-based model	0.378 (0.243–0.531)	0.800 (0.696–0.896)	0.466 (0.324–0.600)	0.657 (0.550–0.754)	0.610 (0.520–0.710)	0.59
Clinical-data-based model	0.733 (0.585–0.860)	0.618 (0.491–0.741)	0.667 (0.543–0.769)	0.707 (0.609–0.806)	0.670 (0.580–0.760)	0.89
Combined model	0.356 (0.227–0.500)	0.782 (0.667–0.879)	0.438 (0.281–0.572)	0.697 (0.589–0.790)	0.590 (0.490–0.690)	Ref
GMIC model						
Internal testing set						
Image-based model	0.640 (0.500–0.767)	0.760 (0.633–0.872)	0.681 (0.571–0.784)	0.833 (0.753–0.908)	0.700 (0.610–0.790)	0.57
Clinical-data-based model	0.860 (0.759–0.952)	0.660 (0.522–0.791)	0.782 (0.690–0.862)	0.790 (0.680–0.873)	0.760 (0.670–0.840)	0.22
Combined model	0.776 (0.681–0.860)	0.800 (0.692–0.907)	0.776 (0.681–0.860)	0.865 (0.785–0.934)	0.780 (0.700–0.850)	Ref
External testing set						
Image-based model	0.644 (0.512–0.780)	0.600 (0.468–0.736)	0.604 (0.483–0.720)	0.674 (0.568–0.777)	0.620 (0.530–0.710)	0.32
Clinical-data-based model	0.800 (0.685–0.913)	0.564 (0.439–0.696)	0.686 (0.571–0.777)	0.705 (0.604–0.804)	0.670 (0.570–0.770)	0.58
Combined model	0.756 (0.625–0.878)	0.618 (0.480–0.746)	0.680 (0.562–0.781)	0.746 (0.650–0.840)	0.680 (0.590–0.770)	Ref

Table 2. Performance of the demodexnet models. CI, confidence interval; GMIC, Globally-aware Multiple Instance Classifier; Ref, reference model; ROC-AUC, area under the receiver operating characteristic curve; SE, stacking ensemble[a]Calculated using the micro-averaged value of each severity class for the given model, using bootstrap resampling ($N = 1000$) of the test dataset[b]The *P*-value from the binomial test measures the difference in performance between the combined model and image- or clinical data-based model in terms of ROC-AUC.

Effect of AI predictions on human responses

Figure 3a and b illustrate the interrater agreement rates among participant responses, AI model predictions, and gold standard labels. In Part I, significant disagreements were observed among participants (Fig. 3a, Fleiss' κ = 0.388 [95% CI, 0.374–0.401]). In contrast, Part II showed a significant improvement in interrater agreement (Fig. 3b, Fleiss' κ = 0.453 [95% CI, 0.439–0.466]). Interestingly, while the interrater agreement rate between the low-experience and expert groups was notably lower than that in other group pairs in Part I, this disparity showed substantial improvement after AI assistance in Part II.

We sought to examine the selection tendencies of human evaluators and AI in successfully predicting high *Demodex* mite density cases, based on their actual clinical diagnoses. Two dermatologists reviewed and labeled the final clinical diagnoses of 100 cases from the internal test set. We then analyzed the distribution of clinical diagnoses among correct answers given by humans, humans with AI assistance, and the AI model alone.

Overall, the AI model showed higher accuracy for high *Demodex* cases (44.0% vs. 23.8%), whereas dermatologists performed slightly better in low *Demodex* cases (34.7% vs.31.5%). Human evaluators tended to select papulopustular rosacea more frequently than the AI model in high *Demodex* cases (58.4% vs. 41.1%), whereas the AI model showed a higher tendency to choose erythematotelangiectatic rosacea than humans (27.5% vs. 16.0%).

In low *Demodex* cases, humans favored erythematotelangiectatic rosacea selections compared with the AI model (10.1% vs. 2.9%), whereas the AI model showed a higher tendency to select atopic dermatitis or acne/folliculitis. Notably, human evaluators assisted by AI demonstrated intermediate values in both overall accuracy and relative distribution of clinical diagnoses, falling between unassisted humans and the AI model alone (Fig. 3c and d and Supplementary Table S3).

Explainable AI models

Feature importance rankings based on absolute SHAP values revealed that the presence of extra-facial skin lesions had the most decisive influence on the model's output. This was followed by age, eosinophil cationic protein, total IgE, eosinophil count, flushing, and positive patch test results (Supplementary Figure S5). Subsequent

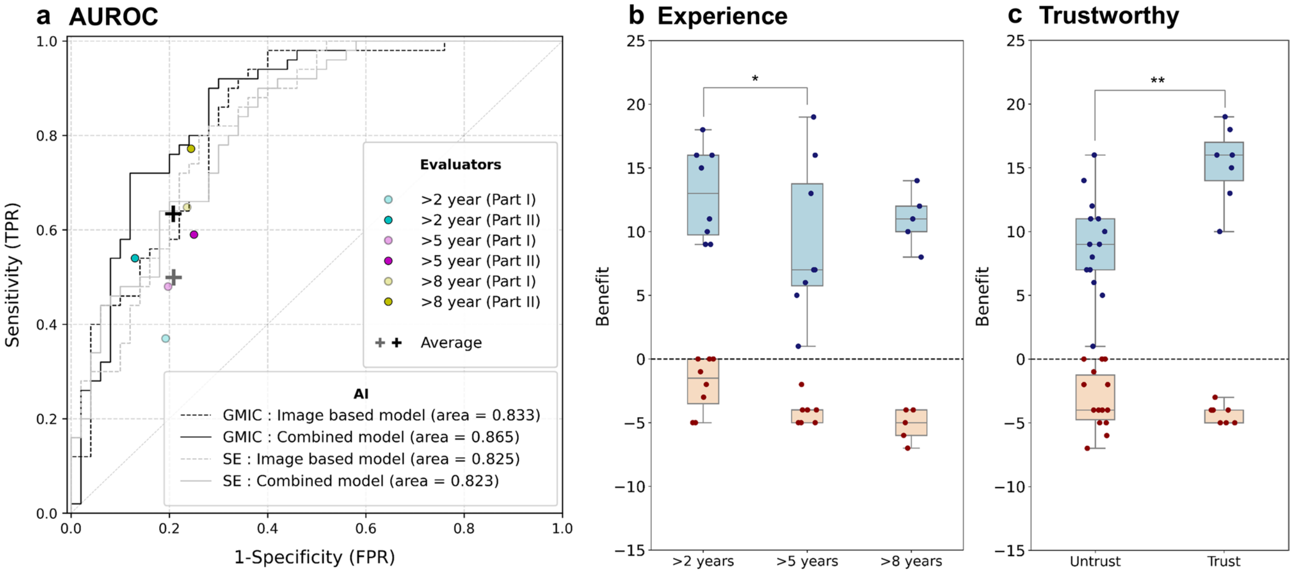

Fig. 2. Performance of DemodexNet and its impact on human evaluators. (**a**) Area under the receiver operating characteristic curve for the DemodexNet models and human evaluators before and after AI assistance. (**b**) Benefit of AI assistance stratified by evaluator experience level. (**c**) Benefit of AI assistance based on evaluators' trust in AI. GMIC, Globally-aware Multiple Instance Classifier; SE, stacking ensemble; TPR, true positive rate; FPR, false positive rate; AI, artificial intelligence. "Benefit" is defined as the change in response from Part 1 to Part 2 that aligns with the response generated by AI, scored as positive when the changed response matches the gold label and negative when it does not. *P<.05, **P<.001 for Mann–Whitney U test.

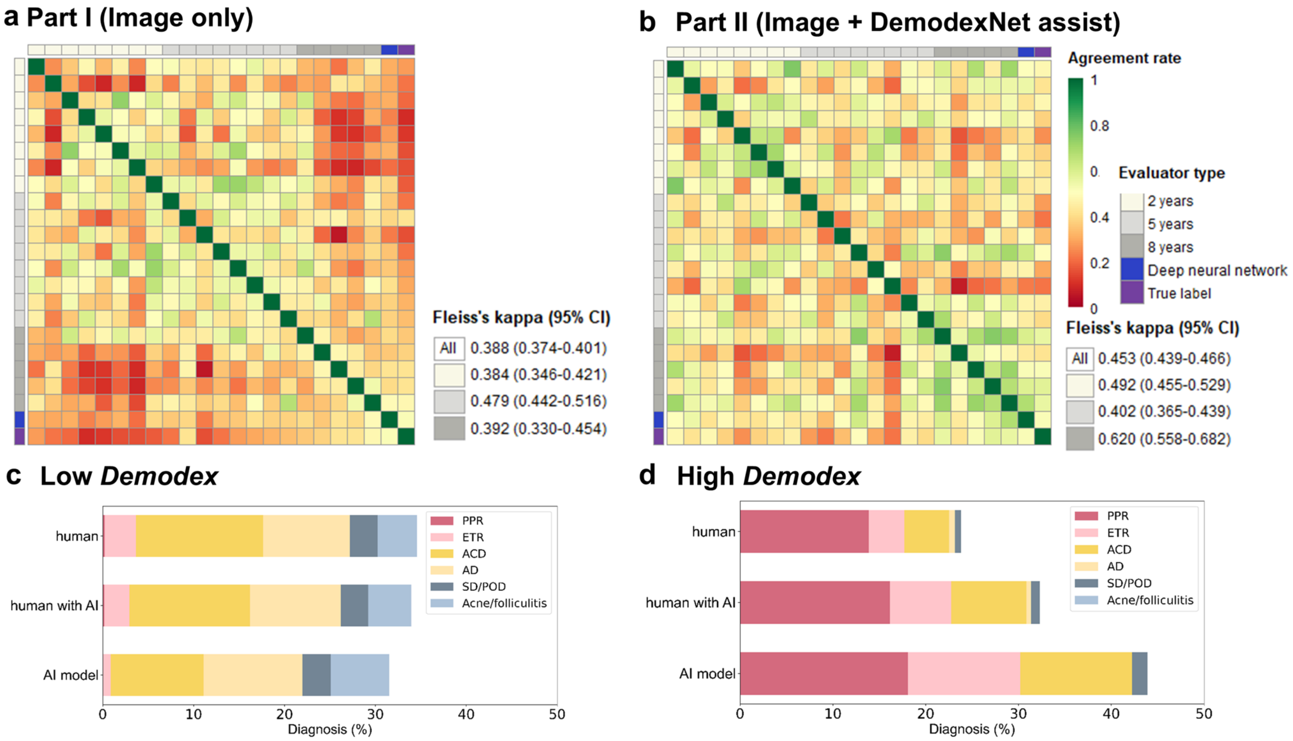

Fig. 3. Impact of DemodexNet assistance on interrater agreement and diagnostic patterns. (**a**) Interrater agreement heatmap for Part I (image only). (**b**) Interrater agreement heatmap for Part II (image + DemodexNet assist). (**c**) Distribution of diagnoses for low *Demodex* cases. (**d**) Distribution of diagnoses for high *Demodex* cases. PPR, papulopustular rosacea; ETR, erythematotelangiectatic rosacea; ACD, allergic contact dermatitis; AD, atopic dermatitis; SD/POD, seborrheic dermatitis/perioral dermatitis.

multivariate logistic regression analysis was performed using two models: a stepwise variable selection model (Model 1) and a model using the Top-7 features from SHAP analysis (Model 2). Both models consistently showed that age (odds ratio [OR]: 1.32, 95% CI: 1.19–1.46 for Model 1; OR: 1.25, 95% CI: 1.12–1.38 for Model 2) and positive patch test results (OR: 1.60, 95% CI: 1.13–2.25 for Model 1; OR: 1.43, 95% CI: 1.01–2.04 for Model 2) were positively correlated with *Demodex* mite density. Conversely, extra-facial skin involvement was negatively correlated with *Demodex* mite density in both models (OR: 0.16, 95% CI: 0.11–0.24 for Model 1; OR: 0.20, 95% CI: 0.13–0.31 for Model 2, Table 3).

Analysis of saliency maps from the GMIC model and Grad-CAM from the SE model revealed that the AI models primarily recognized the central facial region, the leading proliferation site for *Demodex* mites. Furthermore, the models demonstrated the ability to detect individual skin lesions characteristic of demodicosis, such as fine-scaled papules or tiny pustules (Fig. 4; see also Supplementary Figure S6 for additional examples).

Discussion

Various AI models have been developed for inflammatory dermatoses causing facial erythema. Most of these are based on single convolutional neural network models for classification or severity grading of conditions such as acne or rosacea[14–18]. Particularly in individuals with skin of color, diagnosing facial erythema diseases, such as rosacea, based solely on photographs or clinical findings can be challenging[19]. Although *Demodex* mite density testing aids in differential diagnosis, its limited availability prompted us to propose DemodexNet as a complementary diagnostic tool.

In this study, we developed DemodexNet, a deep learning model that predicts *Demodex* mite density in patients with facial erythema using clinical data and photographs. The model demonstrated considerable performance, with ROC-AUC values of 0.823–0.865 in internal testing. When used as a decision-support tool, DemodexNet was associated with an improvement in diagnostic accuracy from 63.7% to 70.6% (*P* <.001) among participating dermatologists. Notably, less experienced dermatologists and those who trusted DemodexNet more benefited the most from AI assistance without increasing errors. The model primarily recognized facial areas and individual lesions characteristic of demodicosis. Additionally, we identified unique clinical features associated with increased *Demodex* mite density.

Accurate prediction of *Demodex* mite density necessitates a model that incorporates both the distribution of erythema throughout the face and the detailed aspects of individual lesions, while also considering clinical features associated with facial erythema[5,8]. Given the complex anatomical landmarks of the face, deriving *Demodex* mite density directly from a whole face image using a single model is challenging[20]. Therefore, we constructed a deep ensemble model based on SE and GMIC, capable of capturing global and local features from multiple facial subregions while incorporating clinical factors. The SE model uses a parallel arrangement of base models for the whole face and local patches, training global and local features in a complementary manner[20,21]. The GMIC model employs flexible localized patches that vary for each face, thus providing an individualized, patient-tailored approach[22,23].

Several studies have investigated the effect of AI assistance on the diagnostic accuracy of clinicians for various skin conditions, including skin cancers, pressure ulcers, and lupus erythematosus[24–28]. Although the present study focused on a different disease with a distinct dataset, the effect of AI assistance on decision-making among survey participants shared similar aspects with previous research. Specifically, raters with less experience or those with higher trust in AI demonstrated greater performance gains from AI-based support[25–27]. Moreover,

Independent variable	Univariate analysis		Multivariate analysis Model 1: stepwise		Multivariate analysis Model 2: SHAP	
	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
Female sex	2.99 (2.08–4.29)	< 0.001	-	-	-	-
Age at the diagnosis, years[a]	1.52 (1.38–1.67)	< 0.001	1.32 (1.19–1.46)	< 0.001	1.25 (1.12–1.38)	< 0.001
Associated symptoms						
Flushing	2.55 (1.91–3.41)	< 0.001	1.38 (0.99–1.92)	0.058	1.33 (0.96–1.86)	0.09
Itching	0.33 (0.24–0.46)	< 0.001	-	-	-	-
Burning/stinging	2.29 (1.66–3.15)	< 0.001	1.34 (0.94–1.91)	0.11	-	-
Edema	2.11 (1.43–3.11)	< 0.001	-	-	-	-
Dry sense	2.79 (1.95–3.99)	< 0.001	-	-	-	-
Positive patch test result	1.28 (0.95–1.73)	0.11	1.60 (1.13–2.25)	0.008	1.43 (1.01–2.04)	0.045
Extra-facial skin involvement	0.12 (0.08–0.18)	< 0.001	0.16 (0.11–0.24)	< 0.001	0.20 (0.13–0.31)	< 0.001
Serum allergy marker value						
ECP (μg/L)[a]	0.71 (0.64–0.80)	< 0.001	-	-	0.87 (0.76–0.98)	0.025
Eosinophil count (cells/μL)[b]	0.60 (0.52–0.68)	< 0.001	-	-	0.88 (0.74–1.04)	0.14
Total IgE (IU/mL)[b]	0.91 (0.88–0.94)	< 0.001	-	-	0.98 (0.95–1.01)	0.11

Table 3. Univariate and multivariate logistic regression analyses of predictive factors associated with high *Demodex* mite density. CI, confidence interval; ECP, eosinophil cationic protein; IgE, immunoglobulin E; OR, odds ratio; SHAP, SHapley Additive exPlanations [a]The original age and serum levels of ECP are divided by 10[b]The original eosinophil count and total serum IgE level are divided by 100

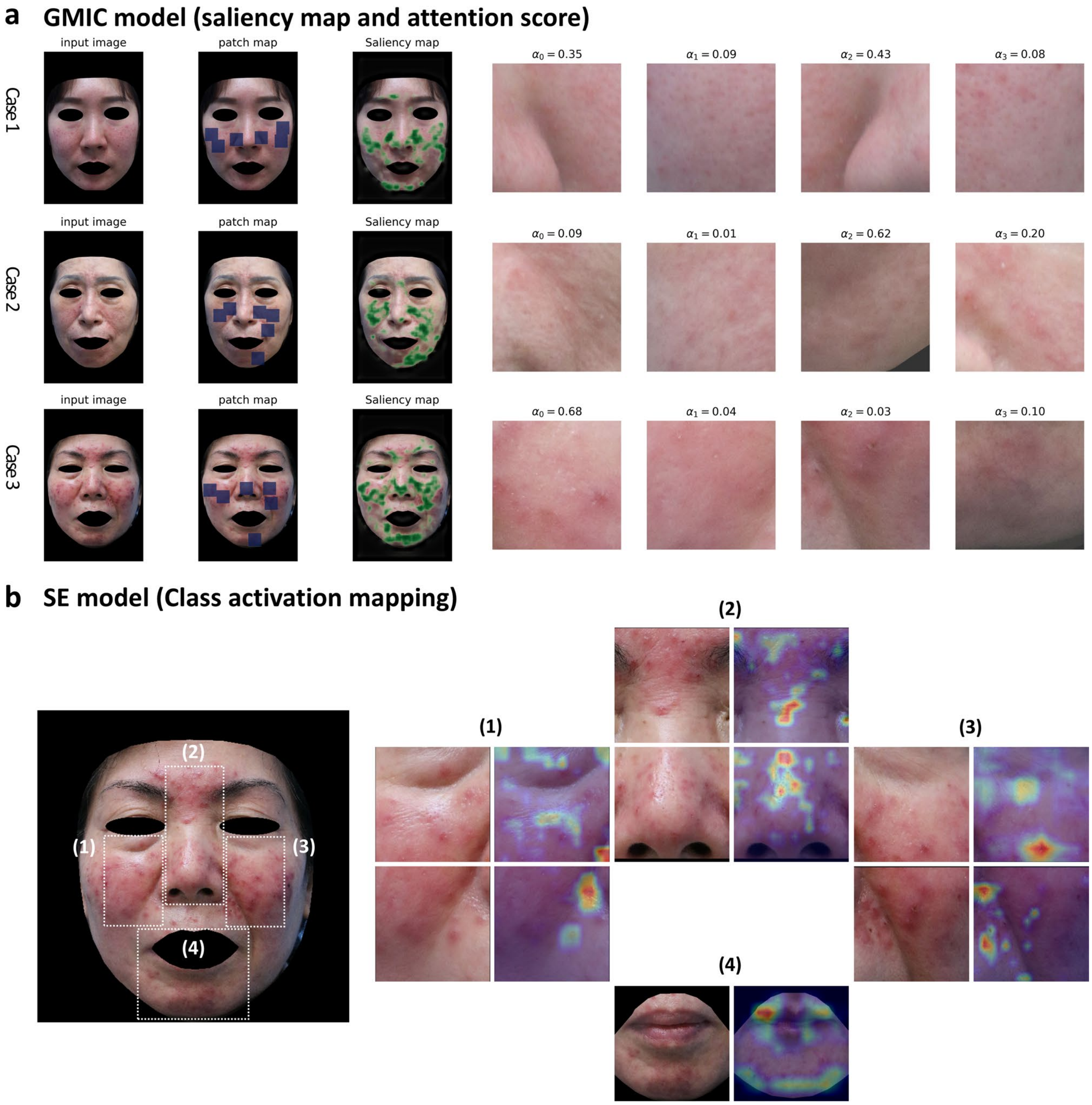

Fig. 4. Visualization of DemodexNet's attention mechanisms. (**a**) GMIC model saliency maps and attention scores for three representative cases. For each case: input image (left), patch map showing regions of interest (center), and saliency map highlighting areas of model focus (right). Attention scores (α) indicate the importance of each patch. (**b**) Class activation mapping of the SE model, showing heat maps of regions contributing to the model decisions. (1) to (4) represent different facial local patches with their corresponding close-up views and activation maps. GMIC, Globally-aware Multiple Instance Classifier; SE, stacking ensemble.

concordance in clinical decisions between diverse groups of experts increased with augmented decision-making[24,27]. Additionally, considering the differences in clinical diagnosis labels between AI and humans, they appear to use distinct diagnostic clues to determine high or low *Demodex* cases. Notably, these differences were mitigated in humans receiving decision support, proving that AI and clinician expertise can complement each other[29].

Our findings suggest that visualization techniques derived from two different model architectures consistently highlight the distribution of *Demodex* mites across the whole face and individual lesions. This attention was particularly pronounced in individuals with typical papulopustular rosacea, known for high *Demodex* mite density (Fig. 3)[30,31]. Interestingly, compared with humans, the AI model more often classified cases as high *Demodex* mites in conditions with atypical facial erythema distribution, such as erythematotelangiectatic rosacea or contact dermatitis (Fig. 3d)[12,32]. This suggests that the decision-making of AI may rely more heavily on recognizing microscopic textures than human assessments[33].

The performance gap between internal and external validation warrants careful consideration. Our analysis revealed two primary factors contributing to this difference. Clinically, the external dataset exhibited substantially different population characteristics, including nearly double the rate of high *Demodex* density (45.0% vs. 24.9%) and varying extrafacial involvement patterns, suggesting distinct clinical phenotypes between institutions. Additionally, inter-examiner variability in the operator-dependent DME method likely introduced differences in ground-truth labeling, despite standardized protocols. From a technical perspective, both our multi-modal architectures (SE and GMIC) may be inherently sensitive to distributional shifts. The SE model's reliance on fixed patches may capture non-discriminative features, whereas GMIC's joint optimization of global, local, and clinical modules with shared gradients could lead to feature misalignment under distribution shift. Although multi-modal fusion effectively captures comprehensive information, it also increases vulnerability to dataset heterogeneity. These findings highlight that the observed performance degradation reflects both real-world clinical variation and architectural limitations, emphasizing the need for domain adaptation strategies in future iterations of the model.

This study has some limitations. First, this work represents a proof-of-concept study with limited generalizability. The data were sourced from two referral hospitals in South Korea, including only a single ethnic group with Fitzpatrick skin types III and IV. This homogeneous population restricts the model's applicability to diverse ethnic backgrounds and other skin phototypes. Future multi-center, multi-ethnic validation studies are essential before clinical deployment in broader populations. Second, our dataset exhibited significant class imbalance, with only 24.9% positive cases in the training set, which may affect model sensitivity and generalization. Third, the external validation cohort was relatively small ($n = 100$), which limited our ability to comprehensively assess model generalizability. Fourth, the DME method used for *Demodex*mite detection is operator dependent[10], which may lead to variations in sensitivity among different examiners. While each institution employed a single experienced examiner (>5 years) using standardized protocols to minimize within-site variability, we lack inter-rater reliability data between institutions, which limits our ability to assess the consistency of our ground truth labels across datasets.

Lastly, the performance of human participants may have been underestimated in our experimental setting, which differed from real-world clinical practice. Participants were provided only frontal facial photographs and limited clinical information, without access to diagnostic tools, such as dermoscopy.

In conclusion, this diagnostic study used clinical data and photographs to develop and evaluate DemodexNet, a deep learning model for predicting *Demodex* mite density in patients with facial erythema. The model showed promising results in predicting *Demodex* mite density and significantly improved the diagnostic accuracy of dermatologists when used as a decision-support tool. The ability of the model to recognize both global facial features and individual lesions characteristic of demodicosis highlights its potential as a valuable aid in clinical practice. Although future studies are needed to validate these results across diverse populations and clinical settings, DemodexNet could potentially aid in clinical evaluation and the management of *Demodex*-related facial erythema, particularly in resource-limited settings or for less experienced clinicians.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to ethical and privacy restrictions. The data contain potentially identifiable facial photographs of patients with dermatological conditions, and public sharing of these images is prohibited by the institutional ethics committees and data protection policies of Yonsei University Health System. Even with deidentification procedures applied, facial features remain inherently identifiable, and patients did not provide consent for public data sharing. However, deidentified clinical data and aggregated statistical results are available from the corresponding author upon reasonable request. The code supporting the findings of this study is openly available at https://github.com/L-YUNNA/demodex_GMIC_pytorch.

Received: 23 September 2025; Accepted: 19 November 2025
Published online: 04 December 2025

References

1. Dessinioti, C. & Antoniou, C. The red face: not always rosacea. *Clin. Dermatol.* **35**, 201–206 (2017).
2. Izikson, L., English, I. I. I., Zirwas, M. J. & J. C. & The Flushing patient: differential diagnosis, workup, and treatment. *J. Am. Acad. Dermatol.* **55**, 193–208 (2006).
3. Haber, R. & El Gemayel, M. Comorbidities in rosacea: a systematic review and update. *J. Am. Acad. Dermatol.* **78**, 786–792 (2018). e788.
4. Olazagasti, J., Lynch, P. & Fazel, N. The great mimickers of rosacea. *Cutis* **94**, 39–45 (2014).
5. Gallo, R. L. et al. Standard classification and pathophysiology of rosacea: the 2017 update by the National rosacea society expert committee. *J. Am. Acad. Dermatol.* **78**, 148–155 (2018).
6. NUTTING, W. B. Hair follicle mites (Acari: Demodicidae) of man. *Int. J. Dermatol.* **15**, 79–98 (1976).
7. Forton, F. & Seys, B. Density of demodex folliculorum in rosacea: a case-control study using standardized skin-surface biopsy. *Br. J. Dermatol.* **128**, 650–659 (1993).
8. Forton, F. Papulopustular rosacea, skin immunity and demodex: pityriasis folliculorum as a missing link. *J. Eur. Acad. Dermatol. Venereol.* **26**, 19–28 (2012).
9. Lee, S. G. et al. Cutaneous neurogenic inflammation mediated by TRPV1–NGF–TRKA pathway activation in rosacea is exacerbated by the presence of demodex mites. *J. Eur. Acad. Dermatol. Venereol.* **37**, 2589–2600 (2023).
10. Aşkın, Ü. & Seçkin, D. Comparison of the two techniques for measurement of the density of demodex folliculorum: standardized skin surface biopsy and direct microscopic examination. *Br. J. Dermatol.* **162**, 1124–1126 (2010).

11. Forton, F. et al. Demodicosis and rosacea: epidemiology and significance in daily dermatologic practice. *J. Am. Acad. Dermatol.* **52**, 74–87 (2005).
12. Forton, F. & De Maertelaer, V. Erythematotelangiectatic rosacea May be associated with a subclinical stage of demodicosis: a case–control study. *Br. J. Dermatol.* **181**, 818–825 (2019).
13. Zhou, J. et al. Pre-trained multimodal large Language model enhances dermatological diagnosis using SkinGPT-4. *Nat. Commun.* **15**, 5649 (2024).
14. Zhao, T., Zhang, H. & Spoelstra, J. A computer vision application for assessing facial acne severity from selfie images. *arXiv preprint arXiv:.07901* (2019). (2019). (1907).
15. Binol, H. et al. Ros-NET: A deep convolutional neural network for automatic identification of rosacea lesions. *Skin. Res. Technol.* **26**, 413–421 (2020).
16. Lim, Z. V. et al. Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin. Res. Technol.* **26**, 187–192 (2020).
17. Yang, Y. et al. Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. *Dermatology Therapy*. **11**, 1239–1248 (2021).
18. Zhao, Z. et al. A novel convolutional neural network for the diagnosis and classification of rosacea: usability study. *JMIR Med. Inf.* **9**, e23415 (2021).
19. Alexis, A. F. et al. Global epidemiology and clinical spectrum of rosacea, highlighting skin of color: review and clinical practice experience. *J. Am. Acad. Dermatol.* **80**, 1722–1729 (2019). e1727.
20. Fan, Y., Lam, J. C. & Li, V. O. (2018) In Artificial neural networks and machine learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, Proceedings, Part I 27. 84–94 (Springer).
21. Almasi, R., Vafaei, A., Kazeminasab, E. & Rabbani, H. Automatic detection of microaneurysms in optical coherence tomography images of retina using convolutional neural networks and transfer learning. *Sci. Rep.* **12**, 13975 (2022).
22. Shamout, F. E. et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *NPJ Digit. Med.* **4**, 80 (2021).
23. Shen, Y. et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med. Image Anal.* **68**, 101908 (2021).
24. Lee, S. et al. Augmented decision-making for acral lentiginous melanoma detection using deep convolutional neural networks. *J. Eur. Acad. Dermatol. Venereol.* **34**, 1842–1850 (2020).
25. Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
26. Han, S. S. et al. Evaluation of artificial intelligence–assisted diagnosis of skin neoplasms: a single-center, paralleled, unmasked, randomized controlled trial. *J. Invest. Dermatol.* **142**, 2353–2362 (2022). e2352.
27. Kim, J. et al. Augmented decision-making in wound care: evaluating the clinical utility of a deep-learning model for pressure injury staging. *Int. J. Med. Inf.* **180**, 105266 (2023).
28. Li, Q. et al. Human-multimodal deep learning collaboration in 'precise'diagnosis of lupus erythematosus subtypes and similar skin diseases. *J Eur. Acad. Dermatol. Venereol.* **38**(12), 2268–2279 (2024).
29. Farzaneh, N., Ansari, S., Lee, E., Ward, K. R. & Sjoding, M. W. Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. *NPJ Digit. Med.* **6**, 62 (2023).
30. Chang, Y. S. & Huang, Y. C. Role of demodex mite infestation in rosacea: a systematic review and meta-analysis. *J. Am. Acad. Dermatol.* **77**, 441–447 (2017). e446.
31. Sattler, E. C., Hoffmann, V. S., Ruzicka, T., Braunmühl, T. & Berking, C. Reflectance confocal microscopy for monitoring the density of demodex mites in patients with rosacea before and after treatment. *Br. J. Dermatol.* **173**, 69–75 (2015).
32. Kim, J. et al. Contact hypersensitivity and demodex mite infestation in patients with rosacea: a retrospective cohort analysis. *Eur. J. Dermatol.* **32**, 716–723 (2022).
33. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).
34. Daneshjou, R. et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. *JAMA Dermatology*. **158**, 90–96 (2022).
35. Huang, H., Hsu, C. & Lee, J. Thumbnail-squeezing method: an effective method for assessing Demodex density in rosacea. (2021).
36. Ilse, M., Tomczak, J. & Welling, M. in *International conference on machine learning.* 2127–2136 (PMLR).
37. Sanchez-Lengeling, B. et al. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:.10685* (2019). (2019). (1910).
38. Selvaraju, R. R. et al. in *Proceedings of the IEEE international conference on computer vision.* 618–626.
39. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv Neural Inf. Process. Syst* **30** (2017).
40. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378 (1971).

Acknowledgements

We thank the board-certified dermatologists and dermatology residents who participated in the user tests. Additionally, we are thankful for the assistance provided by the Digital Health Center, Yonsei University Health System, in data preprocessing and technical support. Also, we are grateful to Prof. Youngdeok Hwang (Paul H. Chook Department of Information Systems and Statistics, Baruch College, The City University of New York, NY, USA) for his expert statistical consultation.

Author contributions

COP and Jihee K had full access to all the data in the study and took responsibility for the data's integrity and the data's accuracy. Jemin K and YNL contributed equally to this work as co-first authors.Concept and design: Jemin K, Jihee K, YNL, and COP. Acquisition, analysis, or interpretation of data: JB, JHL, YSC, and HK.Drafting of the manuscript: Jemin K, YNL, JB, and Jihee K.Critical review of the manuscript for important intellectual content: IO, JIN, and COP.Statistical analysis: Jemin K, YNL, and JB.Obtained funding: Jemin K, Jihee K, and COP.Administrative, technical, or material support: CL and IO.Supervision: Jihee K, COP, and IN.

Funding

This work was supported by faculty research grants from Yonsei University College of Medicine (6-2023-0098 and 6-2024-0056); research grants from the Korean Dermatological Association and Amore Pacific Skin Science 2023; a grant from the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (grant number HI14C1324); and grants from the National Institutes of Health, including grant number RS-2023-00265913, a National Research Foundation of Korea grant funded by the Korea government (MSIT) (grant number RS-2023-00234581, RS-2024-00344273 and RS-2025-16070039), and a new faculty research seed money grant from Yonsei University College of Medicine (grant number 2024-32-0056). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

Declarations

Competing interests

Dr. Oh is currently employed by LG Chem Ltd. However, the company did not have any role in the study design, data collection and analysis, the decision to publish, or the preparation of this manuscript. All other authors declare no financial or non-financial competing interests.

Ethics

This study was approved by the Institutional Review Board of Yonsei University Severance Hospital and Yongin Severance Hospital (approval numbers 4-2023-1008 and 2023-0382-001, respectively), with a waiver of informed consent for the retrospective analysis of clinical data. For patients whose clinical images are presented in this manuscript, separate written informed consent was obtained for publication of their case details.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-29791-9.

Correspondence and requests for materials should be addressed to J.K. or C.O.P.