Development and assessment of an artificial intelligence–based tool for scoring lesion severity in generalized pustular psoriasis clinical images

Alberto Sabater[1], Alfonso Medela[1], Ignacio Hernández Montilla[1], S. Andy Aguilar Robles[2], Taig Mac Carthy[2], Jordi Mollet Sanchez[3], Ofelia Baniandrés[4], Rosa Izu-Belloso[5], Javier Mataix Díaz[6], Jose Manuel Rueda[7], Virginia Sanz-Motilva[7], Gurpreet Singh S. Chowdhry[8], Juan Semeco[8] and Antonio Martorell[2,7]

JID Innovations (2026);**6:**100521 doi:10.1016/j.xjidi.2026.100521

Generalized pustular psoriasis (GPP) is a rare, chronic, systemic inflammatory disease characterized by unpredictable chronic symptoms and periods of flaring. Assessing GPP severity is important for defining treatment goals and evaluating treatment responses, yet the rarity of the disease and the subjectivity of manual evaluation complicate its standardization. To address this, we developed the automatic Generalized Pustular Psoriasis Physician Global Assessment (GPPGA), a digital imaging tool powered by artificial intelligence. On the basis of the validated GPPGA scoring system, this framework rapidly assesses GPP severity in clinical images using a neural network trained on annotated clinical images of GPP, with consistent performance across diverse sexes, ages, and skin tones. Using 2 datasets (V1 and V2, comprising 332 and 4296 images, respectively, with diverse demographics), automatic GPPGA achieved a Cohen's kappa of 0.74 for V1 and 0.82 for V2, whereas dermatologists achieved 0.79 and 0.80, respectively. The accuracy values for automatic GPPGA were 0.62 for V1 and 0.64 for V2 versus 0.70 and 0.65 for dermatologists, respectively. These results support its potential use for assessing lesion severity from images during healthcare visits. Automatic GPPGA offers a reproducible method that can address the subjectivity of manual evaluation and facilitate the standardization of GPP severity assessment.

Keywords: Artificial intelligence, CADx system digital imaging tool, Generalized pustular psoriasis, Generalized Pustular Psoriasis Physician Global Assessment

INTRODUCTION

Generalized pustular psoriasis (GPP) is a rare, chronic, systemic inflammatory disease with an unpredictable and heterogeneous clinical course (Armstrong et al, 2024; Choon et al, 2022; Choon et al, 2024; Fujita et al, 2018; Navarini et al, 2017), characterized by symptoms such as pustules, erythema, and periods of flaring (Choon et al, 2022; Gooderham et al, 2019). GPP is a potentially life-threatening disease, with mortality rates ranging from 2% to 16% owing to severe complications such as multisystem organ failure and sepsis, which require urgent hospital care (Choon et al, 2022; Gooderham et al, 2019). Although the prevalence of GPP is low, it varies considerably across geographical regions and is higher in females than in males, with a mean age of disease onset of 40.9 years (Choon et al, 2014; Prinz et al, 2023).

Accurate assessment of disease severity is important in clinical practice to help physicians define treatment goals for patients with GPP and to evaluate treatment options and responses (Burden et al, 2022, 2023b). Given the unique pathology and clinical features of GPP, several scoring systems for assessing disease severity have been proposed. These include the Generalized Pustular Psoriasis Physician Global Assessment (GPPGA) and the Generalized Pustular Psoriasis Area and Severity Index, which have been developed through modification of scoring systems used for the assessment of plaque psoriasis (the most common form of psoriasis) to capture skin pustulation, a key manifestation of GPP (Burden

[1]*Department of Medical Data Science, Legit.Health, Bilbao, Spain;* [2]*Department of Clinical Endpoint Innovation, Legit.Health, Bilbao, Spain;* [3]*Department of Dermatology, Vall d'Hebron University Hospital, Barcelona, Spain;* [4]*Department of Dermatology, Hospital General Universitario Gregorio Marañón, Madrid, Spain;* [5]*Department of Dermatology, Hospital de Basurto, Bilbao, Spain;* [6]*Department of Dermatology, Hospital Marina Baixa, Alicante, Spain;* [7]*Department of Dermatology, Hospital de Manises, Valencia, Spain; and* [8]*Boehringer Ingelheim International GmbH, Ingelheim am Rhein, Germany*

Correspondence: Alberto Sabater, Department of Medical Data Science, Legit.Health, Bilbao 48001, Spain. E-mail: albertosabater@legit.health

Abbreviations: AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; CI, confidence interval; CK, Cohen's kappa; GPP, generalized pustular psoriasis; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; ICC, intraclass correlation coefficient; MAE, mean absolute error; MoE, mixture of expert; NSL, nonspecific lesion

Received 20 January 2025; revised 25 June 2026; accepted 12 July 2026; accepted manuscript published online XXX; corrected proof published online XXX

Cite this article as: JID Innovations 2026;6:100521

© 2026 The Author(s). Published by Elsevier Inc. on behalf of Society for Investigative Dermatology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

et al, 2023b; Puig et al, 2023). These systems have been used as the basis of clinical endpoints in trials in GPP (Burden et al, 2023b) and in clinical practice (Puig et al, 2023).

The GPPGA comprises 5 grades of severity for erythema, scaling, and pustulation (0 = clear, 1 = almost clear, 2 = mild, 3 = moderate, and 4 = severe) (Burden et al, 2022), and overall severity is calculated by averaging these 3 individual scores (overall score: 0 = clear, >0 and <1.5 = almost clear, 1.5 to <2.5 = mild, 2.5 to <3.5 = moderate, and ≥3.5 = severe). Although the GPPGA provides a tool for assessing GPP severity, it is limited to some extent (similar to many other scoring systems) by inherent subjectivity in its evaluation and the time it requires to complete. Standardized assessment of disease severity is challenging in GPP owing to its heterogeneous nature, including cyclical variations in disease severity that occur over long periods of time (Burden et al, 2022; Kharawala et al, 2020; Puig et al, 2023). These challenges are particularly pronounced for nondermatologist clinicians, such as primary care physicians, whose diagnostic accuracy for dermatological conditions can be as low as 24% (Escalé-Besa et al, 2024), a variability that can impact the timeliness of specialist referrals and initial patient management. There is, therefore, a need to develop automated tools for GPP, similar to previous tools developed to assess the severity of other skin diseases. These tools are often based on modern deep-learning techniques such as convolutional neural networks and computer vision, as is the case for automated tools that have been used to evaluate the severity of atopic dermatitis (Medela et al, 2022; Pan et al, 2020), hidradenitis suppurativa (Hernández Montilla et al, 2023; Wiala et al, 2024), urticaria (Mac Carthy et al, 2024), plaque psoriasis (Huang et al, 2023), and acne (Huynh et al, 2022; Wang et al, 2023).

In this study, we introduce a fast digital imaging tool powered by artificial intelligence, the automatic Generalized Pustular Psoriasis Physician Global Assessment (AGPPGA). On the basis of deep-learning and computer vision techniques, this framework is designed to aid clinicians by providing immediate, objective, and reproducible severity scores for the different visual components of the GPPGA, namely, pustules, erythema, and scaling. This study evaluates AGPPGA's potential as a clinical decision support tool to improve both accuracy and standardization in severity scoring. Specifically, AGPPGA can first support nonspecialists in reducing diagnostic variability and uncertainty and, second, minimize inter and intraobserver inconsistencies among dermatology experts by delivering automated, reproducible assessments.

RESULTS

The AGPPGA and annotators' performance is evaluated on 2 sets of data—V1 and V2—of 332 and 4296 images, respectively, with diverse demographics. Data are split into 4 folds for training and evaluation. Both datasets were annotated by 2 distinct groups of 3 and 4 European-certified dermatologists, respectively, with recognized expertise in psoriasis and GPP. Annotator performance is measured with intraclass correlation coefficient (ICC), accuracy, Cohen's kappa (CK) with quadratic weights metrics, and mean absolute error (MAE); AGPPGA performance is measured with accuracy, CK, and MAE metrics (more details of the data and metrics can be found in the Materials and Methods section).

Annotator performance

Given the subjective nature of the GPPGA, perfect agreement is not always achieved by dermatologists when manually quantifying the different GPPGA components for each image sample (Regier et al, 2013). To measure the level of agreement between our annotators, who were expert dermatologists, we first evaluated the inter-rater reliability by calculating an ICC for the whole datasets. As a result, the V1 dataset achieved an ICC absolute agreement of 0.74 (pustules 0.71, erythema 0.74, scaling 0.78), and the V2 dataset achieved an ICC absolute agreement of 0.90 (pustules 0.87, erythema 0.91, scaling 0.91). These ICC scores, consistent with prior GPPGA-related research (pustules 0.78, erythema 0.82, scaling 0.76) (Burden et al, 2023a), showed good and excellent inter-rater correlation (Cicchetti, 1994). Secondly, we also measured inter-rater reliability with the average pairwise CK statistic, that is, an average measure of the agreement calculated as CK for each pair of annotators. As a result, the V1 dataset achieved a CK score of 0.55 (pustules 0.53, erythema 0.54, scaling 0.56), and the V2 dataset achieved a score of 0.69 (pustules 0.65, erythema 0.71, scaling 0.72). These CK scores reflect good and very good inter-rater correlation, respectively (Regier et al, 2013). Although both datasets exhibited good inter-rater reliability, the V2 dataset showed higher agreement. This is likely attributable to the number of annotators (4 instead of 3) and to utilizing images of large body areas captured in a controlled and standardized clinical setting with higher resolution rather than solely focusing on the affected body region.

Finally, we evaluated the performance (in terms of accuracy and CK) of each annotator with respect to the consensus of the annotators' labels, which, following standard practice in medical artificial intelligence (England and Cheng, 2019), serves as the training and evaluation ground truth for this study. For simplicity, we report these statistics only for the first cross-validation fold. As observed in Figure 1, there was a varying level of performance among annotators across the different GPPGA visual signs for the V1 dataset. Annotator 3 demonstrated lower scores for pustules (0.61) but higher performance for erythema (0.81). Similarly, the MAE for GPPGA scoring also varied among annotators (0.28–0.35). Nevertheless, there was a sufficient degree of agreement among all 3 annotators to allow for the training and evaluation of deep learning–based models for GPPGA visual sign severity assessment. In the case of the V2 dataset, the level of performance differed from that of dataset V1 (Figure 1). First, pustulation demonstrated a significantly lower score (0.61) because its consensus score was calculated on the basis of the maximum score among annotators instead of following the voting strategy. This demonstrates how annotators 2, 3, and 4 failed to detect pustules (barely visible) in these high-resolution images that capture larger body areas (Figure 2d–f), which also led to higher GPPGA MAE scores (0.28, 0.35, and 0.35, respectively). Second, erythema and scaling were associated with higher CK values in this dataset than in the V1 dataset (0.84 vs 0.75 and 0.82 vs 0.81, respectively). This improved consistency is likely attributable

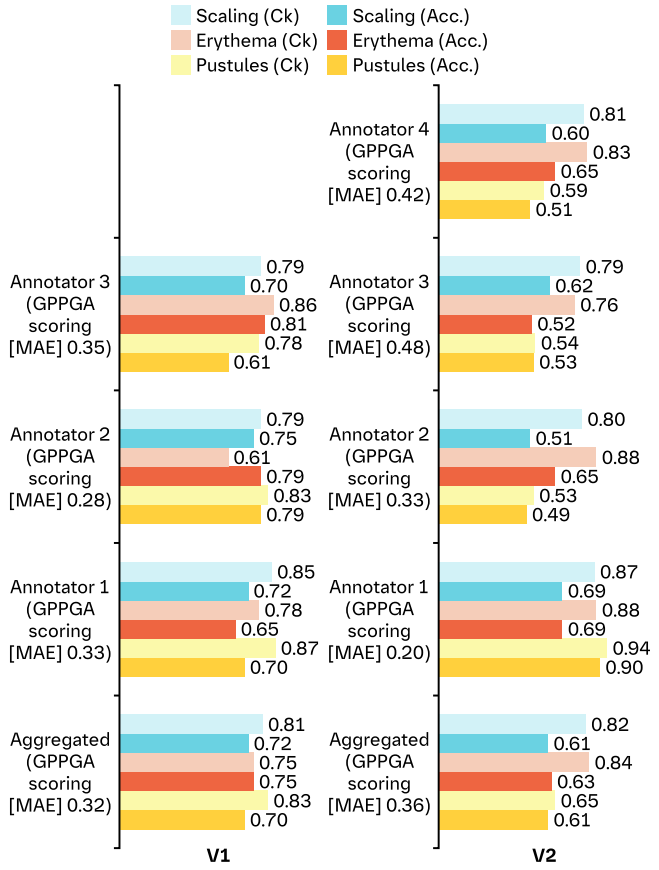

Figure 1. Evaluation of the agreement between annotators in the V1 and V2 datasets with respect to the consensus. Values are reported for the first cross-validation fold. Acc denotes accuracy. CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.

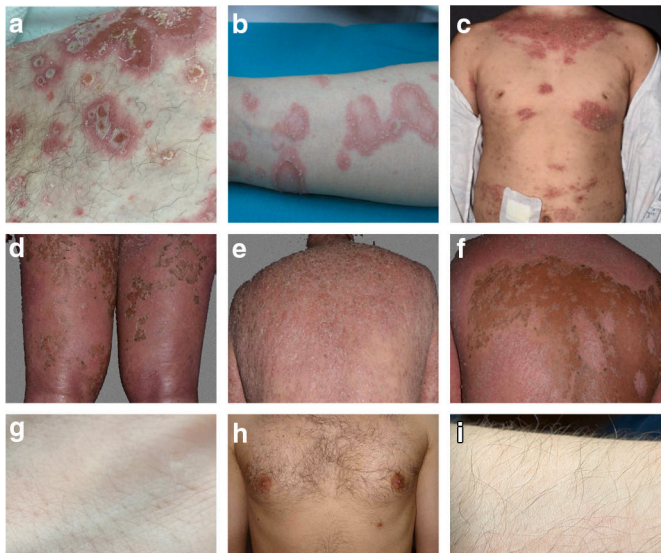

Figure 2. Different datasets built with annotations for each GPPGA component. (**a**–**c**) Images from the V1 dataset supplied by Boehringer Ingelheim Spain. (**d**–**f**) Images from the V2 dataset supplied by Boehringer Ingelheim International GmbH. (**g**–**i**) Images from the nonspecific lesion dataset (source: Legit.Health). (**a**–**c**) Images of generalized pustular psoriasis lesions. (**d**) Image of the lower limbs of a patient with generalized pustular psoriasis. (**e, f**) Images of the back of patients with generalized pustular psoriasis. (**g**–**i**) Images of nonspecific lesions. Images are used in accordance with the applicable data use permissions and ethics approvals. Details are provided in the data and ethics statements submitted with this manuscript. GPPGA, Generalized Pustular Psoriasis Physician Global Assessment.

to the higher standardization of V2 images and the less biased ground-truth labels derived from the larger panel of annotators. Finally, annotator 1 demonstrated a higher score than the other annotators, particularly regarding the identification of pustules (0.90), and annotator 3 demonstrated a lower score than the other annotators for most visual signs.

Overall, variability was observed in the individual annotations, which was mainly caused by subjectivity in the interpretation of the image samples. Nevertheless, using annotations from each annotator makes it possible to account for the high inter-rater variability and facilitates the training and evaluation of the proposed AGPPGA model. Moreover, this inter-rater inconsistency highlights the need for developing an automatic GPPGA scoring tool, which leverages annotators' consensus to measure consistently and reliably the severity of the different GPPGA visual signs.

AGPPGA performance

In the subsequent analysis, we evaluated the performance of our proposed AGPPGA model in assessing the 3 GPPGA visual signs. We then benchmarked this performance against 2 baseline scenarios. The first was a set of 3 single-task models based on the same image encoder used in our AGPPGA model, but which did not leverage multitask learning, resulting in a more cost-intensive solution. The second scenario was the baseline performance achieved by the annotators.

Table 1 shows the performance of the AGPPGA framework on the V1 and V2 validation sets, reported as the mean ± SD across all 4 evaluation folds. The results provide empirical evidence that our mixture of experts (MoE) approach successfully benefits from multitask learning and mitigates negative transfer to improve both efficiency and performance. For instance, on the V1 dataset, AGPPGA showed significant positive transfer for pustulation (CK of 0.76 vs 0.73) and scaling (CK 0.72 vs 0.69), with no significant performance degradation across tasks and datasets. In comparison with expert dermatologists for the visual sign intensity assessment, performance varies across datasets. In the case of the V1 dataset, although AGPPGA achieved competitive performance metrics (accuracy = 0.62 and CK = 0.74), it overall scored below experts (accuracy = 0.70 and CK = 0.79), except for the final GPPGA score calculation (MAE = 0.30 vs 0.32). This is likely attributed to the smaller number of training images in the V1 dataset and its lower inter-rater reliability, which led to higher variability in the annotations. In contrast, in the V2 dataset, AGPPGA achieves comparable or better performance (accuracy = 0.64 and CK = 0.82) than the expert panel (accuracy = 0.65 and CK = 0.80) across most of the visual signs and metrics, with significant improvements in pustulation (CK = 0.71 vs 0.68) and scaling (CK = 0.88 vs 0.86). When analyzing the final GPPGA scoring, AGPPGA achieves a consistently lower MAE in both datasets (MAE 0.30/0.31 vs 0.32/0.33). In terms of reproducibility and stability across data folds, reported SDs

Table 1. Evaluation of the Proposed AGPPGA Framework in the V1 and the V2 Datasets

Evaluator	Pustules		Erythema		Scaling		Aggregated Results		GPPGA Scoring
	Acc.	CK	Acc.	CK	Acc.	CK	Acc.	CK	MAE
V1 dataset									
Single-task[1]	0.57 ± 0.04	0.73 ± 0.04	0.63 ± 0.05	0.74 ± 0.05	0.61 ± 0.07	0.69 ± 0.07	—	—	—
AGPPGA	0.57 ± 0.03	0.76 ± 0.05	0.64 ± 0.05	0.75 ± 0.06	0.64 ± 0.06	0.72 ± 0.04	0.62 ± 0.04	0.74 ± 0.04	0.30 ± 0.01
Annotators	0.64 ± 0.05	0.78 ± 0.04	0.77 ± 0.04	0.80 ± 0.04	0.71 ± 0.02	0.79 ± 0.04	0.70 ± 0.01	0.79 ± 0.02	0.32 ± 0.02
V2 dataset									
Single-task[1]	0.60 ± 0.05	0.75 ± 0.04	0.68 ± 0.01	0.88 ± 0.02	0.70 ± 0.02	0.88 ± 0.02	—	—	—
AGPPGA	0.58 ± 0.04	0.71 ± 0.04	0.66 ± 0.01	0.86 ± 0.02	0.71 ± 0.02	0.88 ± 0.02	0.64 ± 0.02	0.82 ± 0.02	0.31 ± 0.01
Annotators	0.68 ± 0.04	0.68 ± 0.05	0.63 ± 0.01	0.86 ± 0.01	0.64 ± 0.02	0.86 ± 0.03	0.65 ± 0.02	0.80 ± 0.02	0.33 ± 0.02

Abbreviations: Acc., accuracy; AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.

Values are reported as the mean ± SD across the 4 cross-validation folds.

[1]Aggregated and GPPGA scoring results are not calculated for the single-task models because this approach utilizes 3 independent models and, therefore, does not produce parallel visual-sign outputs for an efficient score calculation.

are overall low across visual signs, and there are no significant differences between single-task models, AGPPGA, and experts. Meaning that variations in performance across folds are sourced from the inherent data variability rather than training instability. The overall high agreement with the expert consensus, particularly on the more standardized and patient-split V2 dataset, highlights the potential of AGPPGA as a robust support tool to reduce inter- and intra-rater variability.

In terms of statistical significance, we measured the bootstrap 95% confidence interval (CI) of the difference between the AGPPGA performance and the expert panel in the first cross-validation fold. In the more heterogeneous V1 dataset, AGPPGA is statistically indistinguishable from the expert consensus for both aggregated visual signs (CK: 95% CI = −0.112 to 0.019) and GPPGA (MAE: 95% CI = −0.084 to 0.059). This implies that the model performed within the range of human expertise and was outperformed by only 1 of the 3 individual expert dermatologists in the case of visual signs analysis (CK: 95% CI = −0.1446 to −0.0106). In the case of the most standardized V2 dataset, AGPPGA performance was statistically significantly better, with respect to the expert consensus, than the aggregated expert panel in both aggregated visual signs (CK: 95% CI = 0.002−0.033) and GPPGA (MAE: 95% CI = −0.065 to −0.029) and only outperformed with statistical significance by 1 of the 4 experts (CK: 95% CI = −0.1262 to −0.0939; MAE: 95% CI = 0.0873−0.1290).

In-depth analyses of the first cross-validation fold reveal that when observing the confusion matrices of AGPPGA outcomes (Figure 3), consistency was achieved across each visual sign. Only in severe cases did we observe a subtle drop in performance, likely attributed to their lower presence in the training data. Nevertheless, even in the absence of matched prediction and as previously supported by the CK and MAE statistics, the assessed severity was close to the ground truth, demonstrating the accuracy of the AGPPGA tool.

Performance in relation to patient demographics

Table 2 presents the performance of AGPPGA and the annotators on the V2 dataset, stratified by patient sex. The results indicate a slightly higher performance for the evaluation of female patients than for male patients, likely due to a greater representation of images from female patients in the training data. In addition, although the differences are minor, AGPPGA had higher scores than the annotators across all metrics: accuracy, CK, and MAE.

Table 3 presents performance results for the V2 dataset stratified by patient age. Although performance was overall comparable across all age groups, it was slightly higher for middle-aged subjects (aged 35−49 years) than for young (aged 21−34 years) and elderly (aged 50−66 years) subjects, likely owing to the age distribution in the training dataset.

Table 4 presents the comparative performance of AGPPGA and human annotators across grouped Fitzpatrick skin tones (I−II, III−IV, V−VI) in V1 and V2. Skin tone classification was performed with the Legit.Health software (Legit.Health, 2024) and aggregated for enhanced interpretability. Whereas V1 contained more light-skin images, V2 included a

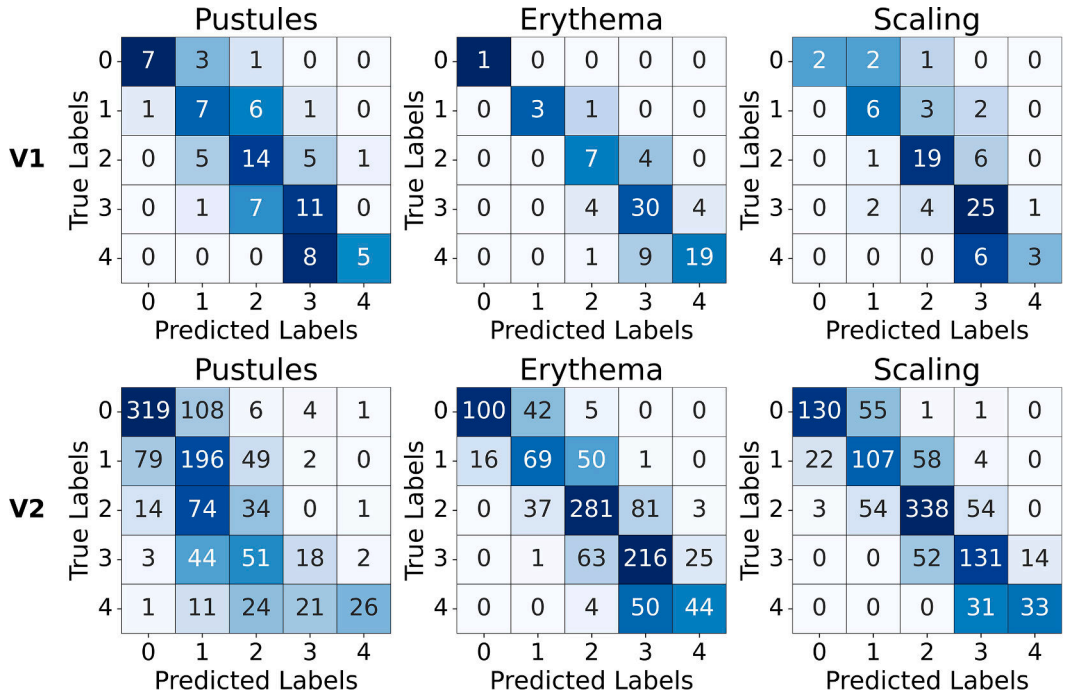

Figure 3. Confusion matrices for the GPPGA visual sign intensity assessment on the V1 and the V2 dataset. Vertical axes refer to the real visual sign score. Horizontal axes refer to the score estimated by AGPPGA. Values are reported for the first cross-validation fold. AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment.

higher proportion of medium-to-dark skin tones. Notably, CK values improved with darker tones for both AGPPGA and experts, likely reflecting the lower proportion of lighter tones during training and the greater visual salience of pustules and scaling on darker skin. In contrast, categorical accuracy showed lower variation across skin tones because it does not account for error magnitude. Comparisons with experts mirror those in Table 1, where annotator performance was stronger in V1, whereas in V2, AGPPGA matched or outperformed experts except in lighter tones. Importantly, AGPPGA yielded consistently lower or equal MAE to that of experts across all skin tones and both datasets.

Pustule consensus analysis

Table 5 reports the performance of the AGPPGA model when using different mathematical consensus criteria. Two strategies were evaluated: a voting-based consensus applied to all visual signs (Voting) and the methodology described in the Materials and Methods (Datasets) section, where pustules in the V2 dataset are assigned the most severe label as the gold standard (Max.), whereas the remaining visual signs follow the voting criterion. The Max. strategy aims to address the difficulty of manually assessing pustule severity in high-dimensional V2 images by prioritizing the most observant annotator, who identified the most subtle pustule clues.

Results in Table 5 suggest that the voting strategy leads to overfitting toward the highly over-represented clear category, resulting in increased accuracy on the V2 dataset, which predominantly contains such labels, without corresponding improvements in CK, which better captures the magnitudes of misclassification. Conversely, the performance with respect to the Max. criterion decreases on the V1 dataset, which exhibits greater diversity in pustule severity labels. This

Table 2. Performance in the V2 Dataset Stratified by Patient Sex

Evaluator	Sex	Number of Patients	Number of Samples	Acc.	CK	GPPGA MAE
AGPPGA	Female	32	760	0.65	0.80	0.29
Annotators				0.63	0.78	0.33
AGPPGA	Male	14	328	0.58	0.74	0.35
Annotators				0.57	0.73	0.41

Abbreviations: Acc., accuracy; AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.

Values are reported for the first cross-validation fold.

Table 3. Performance in the V2 Dataset Stratified by Patient Age

Evaluator	Age, y	Number of Patients	Number of Samples	Acc.	CK	GPPGA MAE
AGPPGA	21–34	11	384	0.59	0.77	0.32
Annotators				0.62	0.81	0.34
AGPPGA	35–49	22	544	0.67	0.81	0.29
Annotators				0.63	0.76	0.35
AGPPGA	50–66	13	160	0.58	0.73	0.35
Annotators				0.54	0.68	0.44

Abbreviations: Acc., accuracy; AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.

Values are reported for the first cross-validation fold.

Table 4. Performance in the V1 and V2 Datasets, Stratified by Skin Tones

		V1 Dataset				V2 Dataset			
			Visual Sign Identification		GPPGA Scoring		Visual Sign Identification		GPPGA Scoring
Evaluator	Fitzpatrick Skin Tone	Number of Samples Train/Evaluation	Acc.	CK	MAE	Number of Samples Train/Evaluation	Acc.	CK	MAE
AGPPGA	I, II	150/55	0.64	0.74	0.30	545/62	0.56	0.68	0.31
Annotators			0.73	0.78	0.31		0.58	0.74	0.34
AGPPGA	III, IV	84/24	0.61	0.78	0.35	1829/615	0.64	0.77	0.30
Annotators			0.69	0.82	0.35		0.62	0.76	0.35
AGPPGA	V, VI	13/4	0.83	—	0.17	834/411	0.61	0.81	0.32
Annotators			0.85	—	0.33		0.61	0.77	0.37

Abbreviations: Acc., accuracy; AGPPGA, automatic Generalized Pustular Psoriasis Physician Global Assessment; CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.
CK cannot be evaluated when samples only belong to an individual category. Values are reported for the first cross-validation fold.

reduction is likely attributable not only to model overfitting but also to label noise introduced by incorrect pustule annotations, which adversely affects learning and makes the model prone to false negatives. These trends are further illustrated in the confusion matrices shown in Figure 4, demonstrating that the proposed Max. strategy promotes a more diverse dataset and improves the identification of subtle pustule clues across varying severity levels. Finally, experimental results demonstrate that a combined strategy, using Max. consensus for pustules and Voting for erythema and scaling, provides the best generalization across heterogeneous imaging conditions, including both highly standardized clinical images and lesion-centered images.

Saliency map visualization

To further validate the reliability of the proposed AGPPGA and the datasets' limitations, we generated saliency maps of the images presented in Figure 2a–f using the well-known gradient-weighted class activation mapping method (Selvaraju et al, 2017). Saliency maps are a widely adopted interpretability tool that highlights the regions of an input image that most strongly influence the model's prediction. As observed in Figure 5, AGPPGA consistently focuses on GPP lesions and their boundaries to assess the degree of pustulation, erythema, and scaling. For instance, in the first image (Figure 5, first row), the model focuses on lighter pixels to evaluate scaling and attends to the transition between healthy skin and erythematous skin to assess erythema. These findings demonstrate that AGPPGA has been effectively trained with sufficient data to prioritize clinically relevant GPP features rather than relying on background, interpatient, or contextual artifacts.

Table 5. Pustule Performance Analysis when Training and Evaluating AGPPGA with Different Mathematical Consensus Criteria on the Pustule Visual Sign

	V1		V2	
Pustules' Mathematical Consensus Criteria	Acc.	CK	Acc.	CK
Voting	0.46	0.69	0.70	0.66
Max.	0.53	0.77	0.54	0.67

Abbreviations: Acc., accuracy; CK, Cohen's kappa.
Values are reported for the first cross-validation fold.

Image encoder selection

Although we based the AGPPGA model on the CAFormer-S image encoder to build the proposed framework, this choice was informed by a comprehensive analysis of other well-known architectures of a similar size. Table 6 shows the performance of these image encoders on the first cross-validation fold. The results indicate that some image encoders, such as ConvFormer-S and EfficientNet-v2-M, produce similar outcomes, whereas others, such as ConvNeXt, do not achieve comparable performance. The superior performance of the CAFormer-S model is attributed to its integration of traditional convolutions with modern attention mechanisms within a framework that effectively leverages the training data.

Performance on the loss function

Table 7 presents the AGPPGA performance across different loss functions. The results highlight that ordinal classification losses enhance performance, particularly in the V2 dataset, surpassing the standard cross-entropy baseline in multiple metrics. Among the evaluated ordinal losses, class distance-weighted cross-entropy (Polat et al, 2022) delivers the most consistent improvements, achieving the highest accuracy (0.64 and 0.63) and CK (0.76 and 0.79) across both datasets and lower MAE (0.31 and 0.32). These findings indicate that class distance-weighted cross-entropy effectively captures ordinal relationships, leading to more precise predictions in visual sign identification and GPPGA scoring.

DISCUSSION

Artificial intelligence has the potential to assist healthcare professionals, including physicians and dermatologists (Huang et al, 2023; Martorell et al, 2022), in diagnosing and assessing the severity of rare conditions such as GPP. The rarity and heterogeneity of GPP present significant challenges in developing a standardized scoring system for assessing disease severity. Despite various proposed severity scores (Havelin and Hampton, 2022; Wang et al, 2023), the GPPGA has been shown to be a specific, robust, and reliable system for grading the severity of pustules, erythema, and scaling (Burden et al, 2023b). However, GPPGA, similar to other

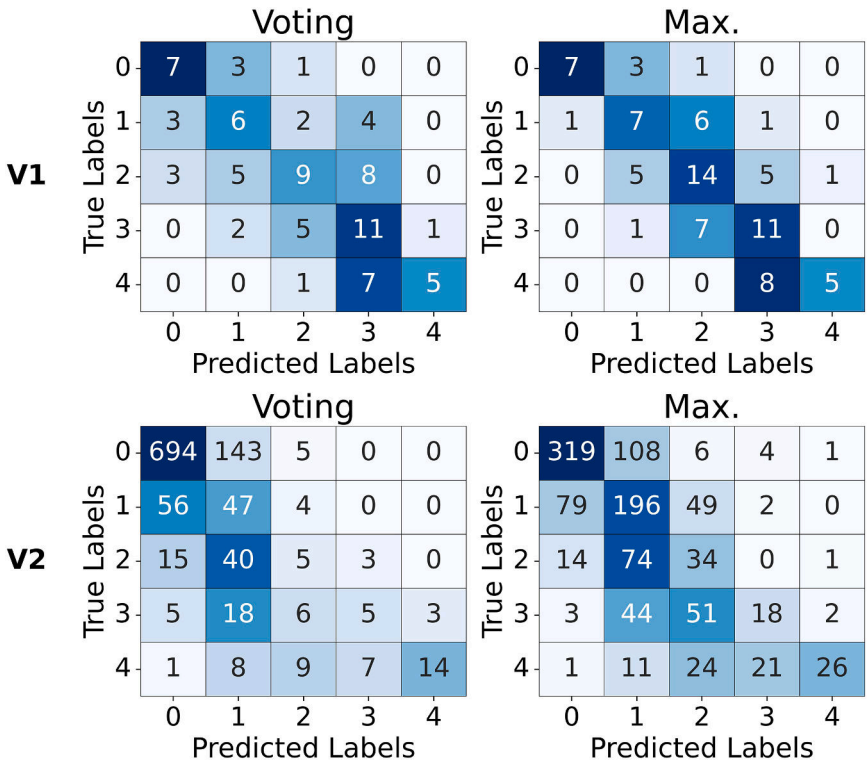

Figure 4. Pustule confusion matrices resulting from training with different pustule consensus criteria. Values are reported for the first cross-validation fold.

scoring systems, suffers from the inherent inter-rater variability of its quantification and the lack of expertise among nondermatologist clinicians. To address this, we designed AGPPGA within the paradigm of augmented intelligence (Kovarik, 2022), not to replace clinical judgment but to assist and elevate human expertise. Its primary role is to provide a reliable, objective, and reproducible baseline to increase the confidence and consistency of clinicians who manage this rare disease.

Our proposed solution addresses these limitations, thanks to a framework based on a deep-learning model that is carefully trained on diverse and expertly labeled datasets to quantify the intensity of pustules, erythema, and scaling. As a result, AGPPGA achieved an overall substantial agreement with expert dermatologists (CK $\geq$ 0.74) and precise GPPGA score calculation (MAE $\leq$ 0.31) across visual signs and a diverse range of sexes, ages, and skin tones. Although the performance with respect to expert dermatologists slightly decreases on the V1 dataset (CK of 0.74 vs 0.79), it performs well on the V2 dataset (CK of 0.82 vs 0.80), which presents better image standardization, careful patient-level data partitioning, and annotations with higher inter-rater reliability. In addition, the reported SDs across the 4 different data folds are comparable to those of the expert panel, showing the potential of AGPPGA to perform reliable assessments and ensure reproducibility. This overall performance is also supported by the statistical significance analysis, which shows that AGPPGA's performance was statistically indistinguishable from the aggregated expert panel on the V1 dataset and statistically superior on the more standardized V2 dataset.

Importantly, the reported metrics (CK >0.74 and MAE <0.31) and confusion matrices illustrate that the vast majority of misclassifications are for adjacent severity categories. This demonstrates that the model fails gracefully, because large-magnitude, clinically dangerous errors (eg, classifying a severe case as clear) are exceptionally rare. This low-risk error profile is a critical feature for building clinician trust and for being a valid support tool in clinical workflows, providing objective and reproducible assessments to guide clinical decisions. This standardization is especially important in rare disease research and multicenter clinical trials, where subjective inter-rater variability can obscure true treatment effects.

It is important to note the distinction between the task performed in this study, where both the human annotators and the AGPPGA model scored clinical images, and a comprehensive in-person clinical assessment, where a doctor would analyze the patient's entire affected area. Moreover, AGPPGA is intended for the severity grading of GPP lesions, not their diagnosis. Consequently, AGPPGA requires either a photograph of the body surface most affected by GPP or multiple images whose individual severity results must be aggregated by the clinician to determine the overall severity on the basis of their clinical judgment. Using images of unrelated pathologies will lead to undesired outcomes. Furthermore, although AGPPGA offers advantages over manual GPPGA assessment, its performance, similar to that of any computer vision model, is highly dependent on the quality of the input images. To ensure optimal precision, images should be captured under good lighting conditions and be free from motion blur. In this context, high-resolution images taken at an appropriate distance are essential, allowing for the identification of small details such as pustules while also preserving surrounding visual cues. Although

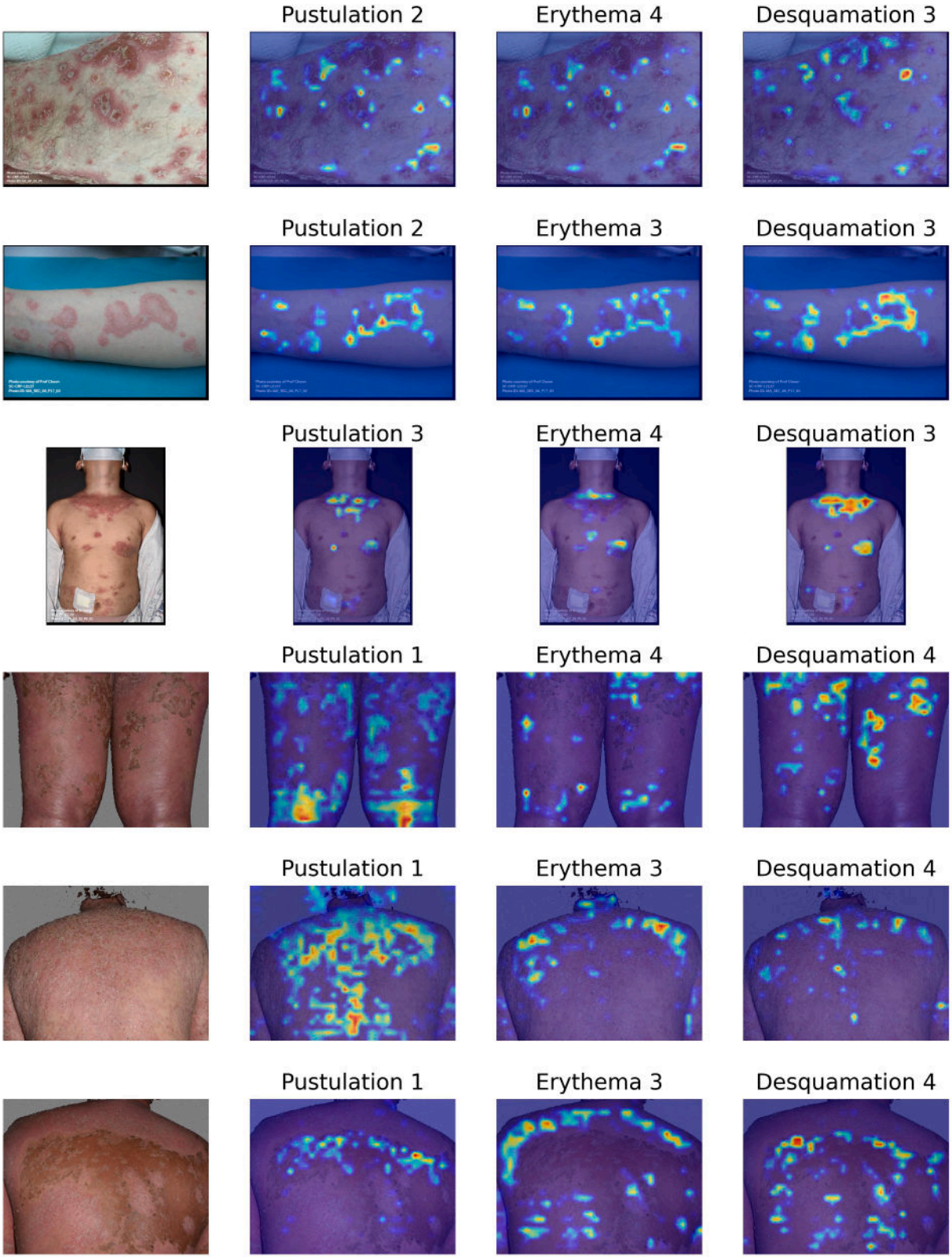

Figure 5. Saliency maps calculated over images from the V1 and V2 datasets. Images are used in accordance with the applicable data use permissions and ethics approvals. Details are provided in the data and ethics statements submitted with this manuscript.

AGPPGA has the potential to facilitate patient self-evaluation, certain images, such as those of the back, would necessitate assistance from a second person. In addition, it must be noted that a comprehensive clinical examination of GPP extends beyond visual analysis alone (eg, oral and genital mucous membranes, skin peeling under

Table 6. Performance of Different Image Encoders to Assess the Severity of the Different GPPGA Components

		V1			V2		
Evaluator	**Image Encoder Parameters[1] (M)**	**Acc.**	**CK**	**GPPGA MAE**	**Acc.**	**CK**	**GPPGA MAE**
ResNet-18 (HuggingFace.com, 2024i)	11.7	0.59	0.72	0.34	0.56	0.74	0.39
ResNet-34 (HuggingFace.com, 2024j)	21.8	0.56	0.73	0.35	0.55	0.75	0.39
ConvNeXt-v2 nano (HuggingFace.com, 2024c)	15.6	0.56	0.68	0.35	0.59	0.76	0.33
ConvNeXt-v2 tiny (HuggingFace.com, 2024d)	28.6	0.56	0.68	0.36	0.59	0.76	0.35
EfficientNet-v2-S (HuggingFace.com, 2024l)	21.5	0.60	0.74	0.29	0.61	0.75	0.35
EfficientNet-v2-M (HuggingFace.com, 2024k)	54.1	0.57	0.73	0.31	0.58	0.75	0.36
MobileNet-v4 Hybrid Medium (HuggingFace.com, 2024h)	11.1	0.52	0.66	0.33	0.57	0.74	0.36
MobileNet-v4 Conv Medium (HuggingFace.com, 2024f)	9.7	0.57	0.69	0.32	0.58	0.76	0.35
MobileNet-v4 Hybrid Large (HuggingFace.com, 2024g)	37.8	0.51	0.61	0.30	0.50	0.68	0.26
MobileNet-v4 Conv Large (HuggingFace.com, 2024e)	32.6	0.61	0.74	0.30	0.58	0.71	0.40
ConvFormer-S (HuggingFace.com, 2024b)	26.8	0.61	0.74	0.32	0.61	0.77	0.34
CAFormer-S (HuggingFace.com, 2024a)	26.3	0.64	0.76	0.31	0.63	0.79	0.32

Abbreviations: Acc., accuracy; CK, Cohen's kappa; GPPGA, Generalized Pustular Psoriasis Physician Global Assessment; MAE, mean absolute error.
Values are reported for the first cross-validation fold.
[1]Based on the timm library that contains nonofficial implementations of these models. The values of parameters may vary slightly from the parameters in the original publications owing to changes in model library dependencies or versions. Parameters from the original publications often included extra layers used for pretraining, which were not required for our study purposes.

pressure, adenopathy). Assessment of GPP severity should also take into consideration medical/drug history; the timing of symptom onset; and accompanying symptoms, for example, skin pain, systemic symptoms, general malaise and physical weakness, cardiac comorbidities, and the patient's level of consciousness.

The performance of the AGPPGA tool is also influenced by the quality of its training data, which is limited by availability and subject to interobserver variability among annotators. Given the rarity of GPP, assembling a diverse dataset is particularly challenging. As a result, the compiled V1 and V2 datasets did not present a uniform distribution of the severity of the individual visual signs, and annotators often struggled to reach a consensus; this was evident in the evaluation of erythema in the V1 dataset and pustules in the V2 dataset. In contrast, the combination of both datasets achieves a good representation of the patients' demographics. Nevertheless, although increased dataset size and reduced interobserver variability would likely enhance the performance of the AGPPGA, the training images curated for this study have already established consistency across different severity grades for each visual sign and patient demographics.

Although the Generalized Pustular Psoriasis Area and Severity Index scoring system also assesses the visual signs of GPP, it includes components relating to body surface area calculation and body part segmentation, which were not considered within the scope of this analysis. Body surface area and segmentation were not considered in the design of the AGPPGA tool, because these features can be assessed more quickly and reliably by the clinician in in-person clinical practice. Finally, owing to the location of this study and the low GPP prevalence, the training and evaluation datasets predominantly featured patients with light and medium skin phototypes (Fitzpatrick I–IV), although other skin tones also presented a relevant proportion of the data. Future research will explore the framework's applicability in clinical practice and with a larger population size with diverse skin demographics.

MATERIALS AND METHODS

GPPGA

The GPPGA is a clinical severity assessment system for GPP (Figure 6). It is a 5-point scale used to assess the severity of 3 different components: pustules, erythema, and scaling.

Table 7. Performance in the V1 and V2 Datasets when Using Different Loss Functions

	V1 Dataset			V2 Dataset		
	Visual Sign Identification		GPPGA Scoring	Visual Sign Identification		GPPGA Scoring
Evaluator	**Acc.**	**CK**	**MAE**	**Acc.**	**CK**	**MAE**
CE (Goodfellow et al, 2016)	0.57	0.71	0.35	0.57	0.73	0.39
EMD (Hou et al, 2016[4])	0.53	0.71	0.36	0.59	0.73	0.37
HO2 (Albuquerque et al, 2021)	0.57	0.73	0.33	0.58	0.75	0.37
CDW-CE (Polat et al, 2022)	0.64	0.76	0.31	0.63	0.79	0.32

Abbreviations: Acc., accuracy; CDW-CE, class distance-weighted cross-entropy; CE, cross-entropy; CK, Cohen's kappa; EMD, earth mover's distance; GPPGA, Generalized Pustular Psoriasis Physician's Global Assessment; HO2, ordinal entropy loss function; MAE, mean absolute error.
Values are reported for the first cross-validation fold.

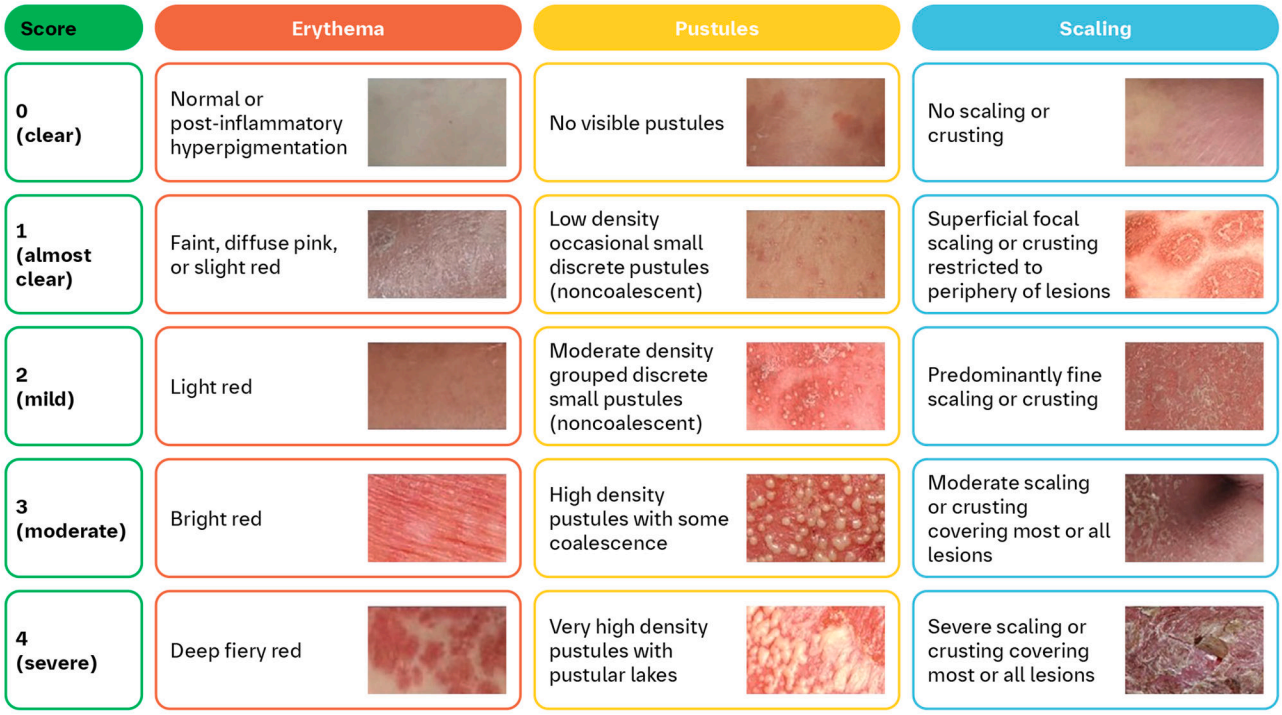

Figure 6. Visualization of the GPPGA scoring system (Burden et al, 2023a)**.** Image was adapted from Burden et al (2023a). Image was published under a CC BY license. GPPGA, Generalized Pustular Psoriasis Physician Global Assessment.

Each component is scored on a scale of 0 to 4: 0 = clear, 1 = almost clear, 2 = mild, 3 = moderate, and 4 = severe. The final GPPGA score is the mean of the 3 component scores (Burden et al, 2023a), defined in the following formula:

$$\text{GPPGA} = \frac{\text{pustules} + \text{erythema} + \text{scaling}}{3}$$

AGPPGA framework

The proposed framework leverages a deep learning–based computer vision model to perform a comprehensive analysis of GPPGA visual signs from photographs (the schematic is presented in Figure 7b). The model architecture integrates an image encoder, a multitask-oriented MoE block, and 3 task-specific classification heads. Each head is responsible for predicting a distinct visual sign (pustulation, erythema, and scaling). This design enables joint training and multitask optimization for the assessment of 3 distinct GPPGA visual signs, ensuring enhanced accuracy and time efficiency in visual analysis. Figure 8 visualizes the artificial intelligence training and inference workflow to enhance the reproducibility of our results. The pseudocode for the model implementation and optimization is publicly available in our open-source GitHub project's repository.

Image encoder

The proposed image encoder model is built using a CAFormer-S model (Yu et al, 2024). This neural network abstracts the well-known Vision Transformer (Dosovitskiy et al, 2020[1]) architecture and replaces the token mixers from the first 2 encoder blocks with depth-wise separable convolutions, while using vanilla self-attention in the remaining 2 bottom encoder blocks. Each of these 4 encoder blocks halves the input height and width dimensions and has output layer dimensions of 64, 128, 320, and 512, respectively. In addition, CAFormer introduces other network optimizations, such as using StarReLU as the activation function for more efficient execution, while maintaining high performance by alleviating possible data distribution shifts.

Multitask learning and MoE

Multitask deep-learning models are efficient solutions because they enable the execution of various tasks, such as GPPGA visual sign intensity assessment, from a single image. This approach eliminates the need for task-specific neural networks, which become cost intensive when used for each task. In addition, these multitask models exhibit superior generalization capabilities by utilizing the information learned across different tasks, reducing the risk of overfitting. Nevertheless, basic multitask methodologies, such as the standard hard parameter sharing (the scheme is presented in Figure 7a), can suffer from negative transfer (Liu et al., 2019). This occurs when the joint optimization of multiple tasks is detrimental to the performance of an individual task compared with a model trained exclusively for that task. This often arises from optimization conflicts, where gradients from different tasks pull the shared model parameters in opposing directions, or from task dominance, where one task monopolizes the shared representation at the expense of others.

We mitigate negative transfer by implementing a multitask architecture based on customized gate control, which utilizes an MoE block to enhance multitask learning (Tang et al, 2020; Gupta et al, 2022[2]). Unlike a rigid shared block, our MoE framework contains both a shared expert, which learns features common to all visual

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv 2020 [preprint]. DOI: abs/2010.11929.

[2] Gupta S, Mukherjee S, Subudhi K, Gonzalez E, Jose D, Awadallah AH, et al. Sparsely activated mixture-of-experts are robust multi-task learners. arXiv preprint arXiv:2204.07689. 2022.

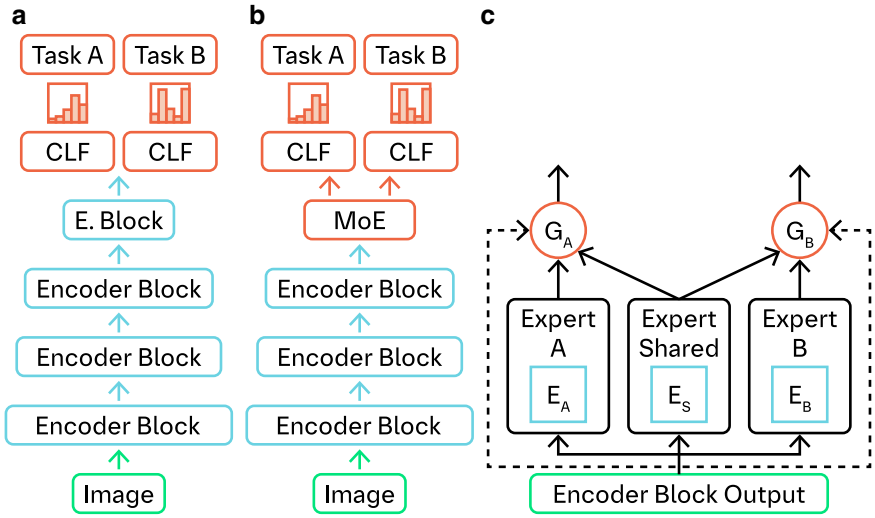

Figure 7. Schematic representations of the model architectures and data workflows to illustrate the structural concepts. (**a**) Baseline hard-parameter sharing architecture: sample tasks A and B are optimized using information from the same image features extracted with an image encoder and task-specific CLFs. (**b**) Proposed model: An MoE block replaces the terminal block of the image encoder to learn shared and task-specific image features, facilitating learning and avoiding negative transfer for the tasks of grading pustules, erythema, and scaling. (**c**) MoE block: Input image features are transformed with the task-specific and shared experts (mirroring the replaced encoder block) and combined accordingly with task-specific gates (G) to serve as input for the final classification heads. Each gate G outputs one softmax weight per expert; the experts' feature maps are combined by weighted sum. CLF denotes classifier, E denotes expert, and G denotes gate. Encoder blocks denote a set of sequential stages (group of layers) of which an image encoder is organized. MoE, mixture of experts.

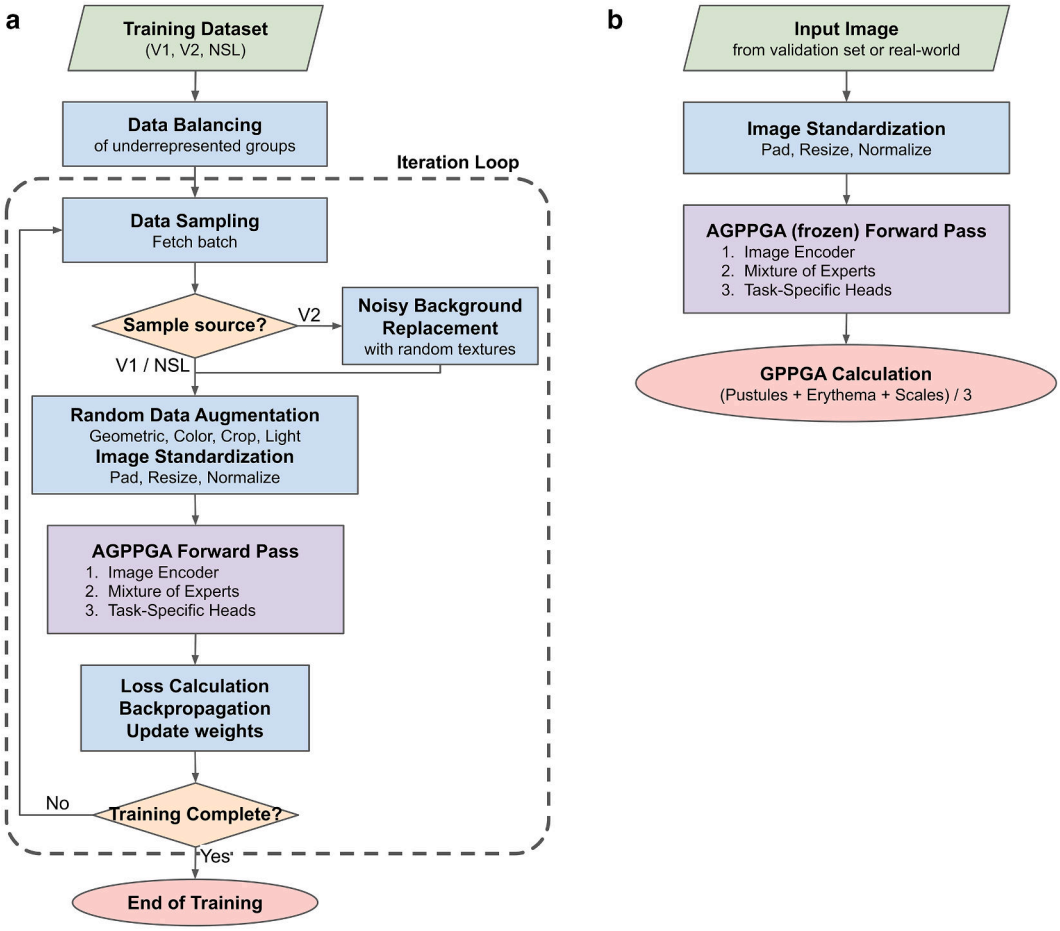

Figure 8. AI workflow for training and inference. **(a)** training and **(b)** inference. AI, artificial intelligence.

signs (facilitating positive transfer), and 3 task-specific experts, which learn the unique, nontransferable features for pustules, erythema, and scaling, respectively. These experts replace and mirror the architecture of the final (fourth) CAFormer-S encoder block (Figure 7b), retaining its LayerNorm normalization and StarReLU activations; consequently, each expert produces feature maps with the 512-channel output dimensionality of that final block. After the single-level customized gate control formulation, that is, the 1-layer instance of porogressive layered extraction (Tang et al, 2020), each task is assigned a dense soft gate (Hwang et al, 2022[3]) that operates on the output of the preceding (third) encoder block, which has a dimensionality of 320 channels. The gate (the scheme is presented in Figure 7c) maps these 320 input channels, through an adaptive average pooling layer, dropout (0.2), and a linear layer with softmax activation, to 2 scalar weights—1 for the task-specific expert and 1 for the shared expert. The task feature map is then obtained as the weighted sum of the two 512-channel expert outputs using these weights. This separation of parameters prevents the learning of one visual sign from interfering with another, directly addressing the root causes of negative transfer.

AGPPGA visual sign heads

The final calculation of GPPGA visual signs is achieved by processing the 3 task-specific outputs with simple classification modules. Each module consists of an adaptive average pooling layer, followed by 2 linear layers (of 512 and 5 dimensionality, respectively), with ReLU and softmax activation functions, respectively, and produces the final classification probabilities, that is, one value for each GPPGA severity score (between 0 and 4).

Optimization

Given that the severity of GPPGA visual signs is quantified using 5 discrete values ranging from 0 to 4, we approach this as an ordinal classification problem rather than a regression task. To leverage the ordinal nature of the classes, we employ class distance-weighted-cross-entropy loss (Polat et al, 2022) for model optimization (Table 7 shows the performance of AGPPGA across different loss functions). This loss function considers the ordinal distance between the ground-truth labels and the predicted labels, facilitating the learning process.

$$CDW - CE = \sum_{i=0}^{N-1} \log\left(1 - y'_i\right) \times |i - c|^{\alpha}$$

where i is the index of each class in the output layer, c is the index of the ground truth ordinal class, y'_i is the probability predicted for class i, and α is a hyperparameter that determines the strength of the class-distance penalty term. Unlike cross-entropy, class distance-weighted cross-entropy evaluates only false class predictions ($|i - c| \neq 0$) using the $log\left(1 - y'_i\right)$ score, which is optimized to be 0. The penalty increases with the ordinal distance from the true class.

During training, we employed different data augmentation techniques (Mumuni and Mumuni, 2022), including geometric, color, crop, and brightness transformations, and standardization with padding, resizing, and normalization. In the case of V2 images, we dynamically replaced their pre-existing Gaussian noise background (only present in V2 images) with textures from the describable textures dataset (Cimpoi et al, 2014) as an additional data augmentation step. This prevented the model from overfitting to the artificial noise and improved generalization to real-world settings. Therefore, note that real-world inference (as in the case of V1 images) does not require any background segmentation or preprocessing step. To improve representation, we upsampled 1.3 times the images from the smaller V1 dataset and balanced V2 images on the basis of the inverse frequency of the severity of pustules, addressing their skewed distribution. In addition, we used dynamic weight average loss weighting (Liu et al, 2019) to ensure a balanced learning strength across each task. We used encoder weights pretrained with ImageNet, with the shared and task-specific experts initialized from the pretrained CAFormer-S final-block weights and then fine tuned; only the gates and classification heads were randomly initialized. We trained with curriculum learning, initially training with a frozen encoder (and unfreezing experts, gating networks, and classification heads) and low-resolution images, progressively increasing image size (in 3 stages of 384, 492, and 600 pixels on the longer side, with 40 epochs per stage), and finally unfreezing the encoder and training at full resolution (600 pixels on the longer image side and 120 epochs). This ensured a solid base of learning, avoided the necessity of larger annotated datasets for fine tuning, and eased model generalization and convergence. Larger images were crucial for detecting small pustules. The training spanned 240 epochs with a batch size of 40. The final model weights were the ones that optimized the whole dataset's aggregated CK in the validation set. For optimization, we also used dropout for regularization (0.2), mixed precision to reduce resource consumption, the Adam optimizer with a 0.0005 learning rate, halved for encoder layers, a weight decay of 0.001, and 5 warmup epochs. These strategies ensured stable convergence and high performance. A single training run took approximately 4–7 hours to complete on an NVIDIA RTX 6000 GPU, depending on the specified architecture and hyperparameters. The complete training pseudocode and detailed training and validation plots have been uploaded to the project's GitHub repository to assist researchers in reconstruction efforts. The model was trained using Python 3.11 and the libraries PyTorch 2.2.1, timm 1.0.7, and albumentations 1.4.2.

Datasets

For the training and evaluation of the proposed framework, 3 different datasets were built, with annotations for each of the GPPGA components (pustules, erythema, and scaling).

The V1 dataset contains 332 images (Figure 2a–c) from patients with GPP, of which 101 samples came from the EFFISAYIL 1 clinical trial (Bachelez et al, 2021), which was proprietary to Boehringer Ingelheim at that time, and the others were extracted from public and open dermatology atlases: DermIS (DermIS, 2024), DermNet and DermNet NZ (DermNet, 2024), Danderm (DANDerm, 2024), DermQuest (DermQuest, 2026), and the DermatologyAtlas (Dermatology Atlas, 2024). Images from this dataset predominantly represent light skin tones and capture the lesion of interest solely, excluding the rest of the patient's body. Patient identifiers are not available in this subset.

The V2 dataset contains 4296 images (Figure 2d–f) from 46 different patients with GPP from the EFFISAYIL 2 clinical trial (Morita et al, 2023), which was proprietary to Boehringer Ingelheim at that time, taken at different times. These images capture large body areas (ie, images of the upper/lower trunk and legs, from the front/back of

[3] Hwang C, Cui W, Xiong Y, Yang Z, Liu Z, Hu H, et al. Tutel: adaptive mixture-of-experts at scale. arXiv 2022 [preprint]. DOI: abs/2206.03382.
[4] Hou L, Yu CP, Samaras D. Squared earth mover's distance-based loss for training deep neural networks. arXiv 2016 [preprint]. DOI: 1611.05916.

the body) using very high-resolution images. Consequently, analyzing minor details such as pustules, which can often be confused with scaling, is challenging with such images. The V2 images were originally taken with a blue background; prior to our acquisition, this background was removed by the original authors using color thresholding and replaced with Gaussian noise. The resulting Gaussian background that we work with can be visualized in Figure 2d–f. The patients predominantly represent medium-to-dark skin tones and are properly identified to stratify the dataset partition.

The nonspecific lesion (NSL) dataset (Figure 2g–i) contains 50 images generated by cropping 29 larger images of healthy skin or NSL images. These images include some nonidentifiable images of Legit.Health employees taken specifically for this work with their written informed consent and others compiled from trackable open Internet images. GPPGA components for this dataset are set to zero (ie, no scaling or crusting, no visible pustules, and normal pigmentation of the skin). This set of data, containing images of healthy skin, was intended to anchor the clear (GPPGA = 0) baseline and increase performance in the assessment of lesions with lower severity. Patient identifiers are not available owing to the rare recurrence of subjects. Before the onset of the study, volunteers received information about the aim of the study and were able to ask any related questions.

Following the recommended practice for clinical data annotation (England and Cheng, 2019), the images in V1 and V2 were annotated by expert dermatologists, whose consensus was used as the ground truth for model training and annotator and model evaluation. This expert-based standard practice aims to reduce individuals specific biases, resulting in a more robust and reliable reference standard that represents the collective clinical judgment of the field. Particularly, the datasets were annotated by European-certified dermatologists with recognized expertise in psoriasis and GPP. The V1 dataset had 3 annotators (JM, OB, and RIB), and the V2 dataset had 4 annotators (JMD, JMR, JMS, and VS). For each image, the GPPGA visual sign intensity scores were determined through consensus among the annotators. Specifically, the final label was either the score that most annotators agreed on or the average of their annotations if no majority was reached. An exception was made for the description of pustules in the V2 dataset. Owing to the difficulty in localizing pustules in V2 high-resolution images without zooming in and the tendency for mislabeling, we opted to use the maximum score among the votes. This strategy aims to prioritize the detection of subtle lesions that might be missed by some annotators (false negatives) but identified by those conducting a more thorough inspection, also allowing for more balanced data labeling. The validity of this approach was empirically demonstrated in the Results section (see Pustule consensus analysis, Table 5, and Figure 4). The distribution of the GPPGA severity grades for the V1 and V2 datasets is shown in Figure 9. In the case of NSL images, the visual sign scores were set to 0, because these images do not exhibit GPP symptoms.

To ensure reproducibility and training stability, we implemented a 4-fold cross-validation learning strategy. The V1, V2, and NSL datasets were partitioned into 4 folds, with each fold using 75% of the data for training and a mutually exclusive 25% for evaluation. For the V2 dataset, where unique patient identifiers were available, the split was performed to ensure strict patient-level partitions across all folds. For the V1 dataset, where most images come from public sources without patient identifiers, patient-level separation could not be completely guaranteed.

Distribution of severity within training and validation datasets

Figure 9 illustrates the distribution of severity grades for pustulation, erythema, and scaling (0 = clear, 1 = almost clear, 2 = mild, 3 = moderate, and 4 = severe) in the training and validation sets from the first data fold and from the V1 and V2 datasets. These grades were assigned on the basis of consensus among the annotators. Although the training and validation sets within the same dataset exhibited similar severity distributions, there were differences between datasets and between visual signs. For example, pustulation in V1 was primarily mild to moderate (grades 1–3), whereas V2 contained mostly mild cases (grades 1–2). Scaling and erythema skewed toward more severe cases in V1 but were more uniform in V2. Overall, V1 had fewer images of low-grade severity, whereas V2 had fewer images of high-grade severity. Despite this imbalance, the distinct characteristics of both datasets provided complementary information for training and validating the AGPPGA tool, enabling it to learn meaningful patterns across all visual signs and severity grades.

Evaluation metrics

The assessment of GPPGA severity is an ordinal classification task, where the 5 outcome classes (0 = clear to 4 = severe) have a natural order. For such tasks, evaluation metrics must be sensitive to the magnitude of a prediction error, because larger errors in severity grading have greater clinical significance. We therefore selected a

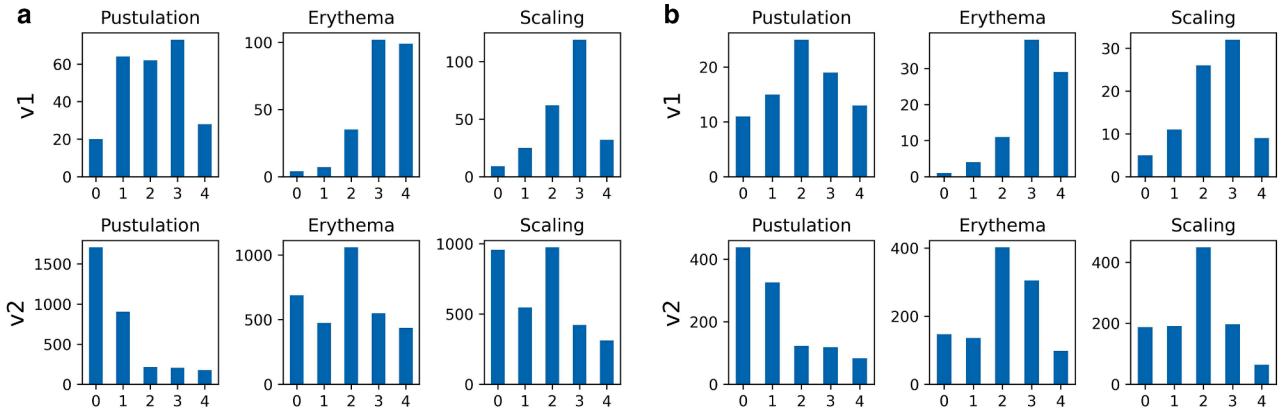

Figure 9. Distribution of severity grades within V1 and V2 datasets for each GPP visual sign. (**a**) Includes the data distribution of the train dataset. (**b**) Includes the data distribution of the validation dataset. Values are reported for the first cross-validation fold. GPP, generalized pustular psoriasis.

comprehensive set of metrics to fully characterize performance. For a general overview, we calculated categorical accuracy, which is not sensitive to error magnitude but is easy to interpret. To analyze distance-sensitive errors, we calculated quadratically weighted CK, which penalizes large errors more heavily than adjacent-class errors, and MAE, which directly quantifies the average error magnitude. We used cross-validation to assess the overall performance and its stability across data splits. Table 1 therefore reports the mean ± SD across the 4 evaluation folds. All subsequent, more granular analyses (Tables 2–7; Figures 1, 3–5, and 9; and the bootstrap CIs) are reported on a single representative fold, the first cross-validation fold, and are not aggregated across folds. The low SDs in Table 1 confirm that performance is stable across folds, so a single fold is representative for these granular analyses, keeping the number of trained models and the analysis tractable.

To determine whether the observed performance differences between the AGPPGA model and the dermatologists were statistically meaningful between the model and both the aggregated expert error and each individual expert, we performed nonparametric bootstrap hypothesis testing for statistical significance. We generated 5000 bootstrap resamples with replacement from the validation sets. For each resample, we calculated the difference (AGPPGA − expert criteria) in the metrics (aggregated visual sign CK and GPPGA MAE) between the model and the expert criteria. Statistical significance was determined by constructing 95% percentile-based CIs; intervals that do not contain zero indicate a statistically significant difference at the $P < .05$ level.

In addition, confusion matrices were used to visually analyze the distribution and magnitude of these errors. Because NSL images lack visual signs of GPP, performance results are exclusively reported for the V1 and V2 datasets.

Performance results are reported separately for each dataset to transparently demonstrate the model's robustness across distinct data modalities (realistic lesion-centered vs highly standardized large captures) and to avoid skewing aggregate metrics toward the significantly larger V2 dataset.

ETHICS STATEMENT

For public dermatology atlases, images were sourced from publicly available dermatology atlases. These images were used in accordance with the copyright policies and ethical guidelines of their respective source platforms. For clinical trial data, images were sourced from the EFFISAYIL 1 (NCT03782792) and EFFISAYIL 2 (NCT04399837) clinical trials. Both trials were conducted in strict accordance with the Declaration of Helsinki, the International Council for Harmonisation Good Clinical Practice guidelines, and applicable local regulations (eg, EU Regulation Number 536/2014, Japanese Good Clinical Practice regulations). The study protocols were approved by the institutional review boards or ethics committees at each participating site. All patients provided written informed consent for trial participation and the use of their images. Confidentiality agreements were in place between the authors and the study sponsor at the time, Boehringer Ingelheim. More detailed information regarding data compilation, sharing policies, review board approvals, and ethical standards can be found in the original source publications for these trials (Bachelez et al, 2021; Morita et al, 2023). Nonspecific lesion dataset consists of images of healthy skin or nonspecific lesions. Internet-sourced subset consists of images compiled from open internet sources and were used in accordance with their respective licensing terms. For consented subset, images were obtained from volunteers (including Legit.Health employees) specifically for this study. Formal written informed consent was obtained from all participants. These images have been open sourced and are available in the project's GitHub repository. Volunteers were informed about the study's aim and what they were required for. Before signing the informed consent, they could ask the researchers any related questions. The informed consent was developed according to the General Data Protection Regulation and in accordance with the Good Clinical Practice guidelines and the Declaration of Helsinki. For institutional approval, owing to the retrospective nature of this study, which used either previously approved clinical trial data or open-source/nonidentifiable images, no separate or additional ethics committee approval was required for this specific computational analysis. Additional details regarding data provenance, ethical compliance, and code availability can be found in the project's GitHub repository: https://github.com/Legit-Health/AGPPGA (https://doi.org/10.5281/zenodo.17975533).

DATA AVAILABILITY STATEMENT

For publicly available datasets, images were sourced from public repositories. A list of URLs for the publicly sourced images used in this study is available in the project's GitHub repository. These images belong to the following datasets: Dermatology Atlas: http://www.atlasdermatologico.com.br/, Danderm: http://danderm-pdv.is.kkh.dk/, DermNet NZ: https://dermnetnz.org/, DermNet: https://dermnet.com/ (available at https://www.kaggle.com/datasets/shubhamgoel27/dermnet/), DermQuest: http://dermquest.com/ (unavailable as of February 2, 2026 therefore images used from this resource are available from the corresponding author), and DermIS: https://www.dermis.net/dermisroot/en/home/index.htm (https://doi.org/10.1111/j.1346-8138.1998.tb02505.x). For clinical trial datasets, private patient images were sourced from the following clinical trials: EFFISAYIL 1 (NCT03782792): "Trial of Spesolimab for Generalized Pustular Psoriasis" (DOI: 10.1056/NEJMoa2111563) and EFFISAYIL 2 (NCT04399837): "Efficacy and safety of subcutaneous spesolimab for the prevention of generalised pustular psoriasis flares (Effisayil 2): an international, multicentre, randomised, placebo-controlled trial" (DOI: 10.1016/S0140-6736(23)01378-8). These images were obtained and used in strict accordance with the ethical approvals, informed consent protocols, and data-sharing specifications detailed in their respective primary publications (Bachelez et al, 2021; Morita et al, 2023). For nonspecific lesion dataset, the dataset's original images of normal subjects are available in the project's GitHub repository. The consent form for the nonspecific lesion dataset was generated and managed internally by Legit.Health to adhere to the General Data Protection Regulation and data privacy standards for nonclinical research. This dataset consists of 2 subsets. The first is internet-sourced subset, whose detailed list of the original source URLs for images sourced from the internet is provided in the repository. These images were used in accordance with their respective source policies and the relevant regulations. The other is consented subset. These images underwent a thorough anonymization process, ensuring that the subjects were nonidentifiable. They were all nonclinical images collected with formal written informed consent from all participants. These images are publicly available in the repository.

For software availability, the automatic Generalized Pustular Psoriasis Physician Global Assessment software code is proprietary to Legit.Health and is not available for public distribution. Nevertheless, to ensure scientific reproducibility, detailed model and training pseudocode scripts, along with training and validation plots, are publicly available in the GitHub project repository. In addition, all methods, model architectures, and data-processing steps used to derive the results are fully reported in the manuscript and Supplementary Materials to ensure reproducibility.

For repository access, further details on data sources, image indexes, pseudocode, and code availability can be found in the GitHub repository (https://github.com/Legit-Health/AGPPGA) (https://doi.org/10.5281/zenodo.17975533).

ORCIDs

Alberto Sabater: http://orcid.org/0000-0002-1156-8617
Alfonso Medela: http://orcid.org/0000-0001-5859-5439
Ignacio Hernández Montilla: http://orcid.org/0000-0003-0356-6619
S. Andy Aguilar Robles: http://orcid.org/0000-0003-0618-6179
Taig Mac Carthy: http://orcid.org/0000-0001-5583-5273
Jordi Mollet Sanchez: http://orcid.org/0009-0008-3880-3729
Ofelia Baniandrés: http://orcid.org/0000-0001-8183-3941
Rosa Izu-Belloso: http://orcid.org/0000-0001-6159-6395
Javier Mataix Díaz: http://orcid.org/0000-0001-5749-1160
Jose Manuel Rueda: http://orcid.org/0009-0004-1394-960X
Virginia Sanz-Motilva: http://orcid.org/0000-0002-0610-8501
Gurpreet Singh S. Chowdhry: http://orcid.org/0009-0007-7884-9337
Juan Semeco: http://orcid.org/0009-0008-4569-3425
Antonio Martorell: http://orcid.org/0000-0003-1378-1590

CONFLICT OF INTEREST

AS, AMe, IHM, AA, and TMC are employees of Legit.Health and have collaborated with Boehringer Ingelheim, which funded this study. OB has acted as a consultant and/or speaker for Boehringer-Ingelheim. RIB has collaborated with Boehringer Ingelheim as an advisory board participant and conference speaker. JMS, OB, RIB, JMD, JMR, and VS have received fees for

their participation in this project. GSC and JS are employees of Boehringer Ingelheim. AMa has received honoraria and/or travel grants and/or has acted as an advisory board member for Novartis, AbbVie, Janssen Cilag, UCB, Lilly, LEO Pharma, L'Oreal, Sanofi, Boehringer Ingelheim, Almirall, Bristol Myers Squibb, and Amgen. He has also worked as a principal investigator in clinical trials supported by AbbVie, UCB, Janssen, Bristol Myers Squibb, Lilly, Galderma, Sanofi, and Novartis.

ACKNOWLEDGMENTS

Sara Sabeh (Nucleus Global, London, United Kingdom) provided editorial and formatting support, which was funded by Boehringer Ingelheim.

AUTHOR CONTRIBUTIONS

Conceptualization: AMe, TMC, AMa, AS; Data Curation: IHM, AS, GSC, OB; Formal Analysis: OB, AS, AMa; Funding Acquisition: AA, GSC; Investigation: VS-M, JMD, OB, JMR, AS, AMa; Methodology: AS, GSC, TMC; Project Administration: AS, AMe, AA; Software: AS; Supervision: IHM, AS, AMe, AA, AMa; Validation: JMD, AS, OB, TMC, JMS, RI-B; Visualization: VS-M, JMD, TMC, JMR, AS, OB; Writing – Original Draft Preparation: AS, AMe; Writing – Review and Editing: IHM, AMe, TMC, AG, JMS, RI-B, JMD, JMR, VS-M, GSC, JS, AS, AMa, OB

DECLARATION OF GENERATIVE ARTIFICIAL INTELLIGENCE (AI) OR LARGE LANGUAGE MODELS (LLMs)

This research involved the use of a digital imaging tool powered by artificial intelligence (Legit.Health). The authors declare no use of artificial intelligence/large language model in the preparation of this article.

REFERENCES

Albuquerque T, Cruz R, Cardoso JS. Ordinal losses for classification of cervical cancer risk. PeerJ Comput Sci 2021;7:e457.

Armstrong AW, Elston CA, Elewski BE, Ferris LK, Gottlieb AB, Lebwohl MG, et al. Generalized pustular psoriasis: a consensus statement from the National Psoriasis Foundation. J Am Acad Dermatol 2024;90:727–30.

Bachelez H, Choon SE, Marrakchi S, Burden AD, Tsai TF, Morita A, et al. Trial of spesolimab for generalized pustular psoriasis. N Engl J Med 2021;385:2431–40.

Burden AD, Bachelez H, Choon SE, Marrakchi S, Tsai TF, Turki H, et al. The generalized pustular psoriasis Physician global assessment (GPPGA) score: online assessment and validation study of a specific measure of GPP disease activity. Br J Dermatol 2023a;189:138–40.

Burden AD, Bissonnette R, Lebwohl MG, Gloede T, Anatchkova M, Budhiarso I, et al. Psychometric validation of the generalized pustular psoriasis physician global assessment (GPPGA) and generalized pustular psoriasis area and severity index (GPPASI). J Eur Acad Dermatol Venereol 2023b;37:1327–35.

Burden AD, Choon SE, Gottlieb AB, Navarini AA, Warren RB. Clinical disease measures in generalized pustular psoriasis. Am J Clin Dermatol 2022;23(Suppl. 1):39–50.

Choon SE, Lai NM, Mohammad NA, Nanu NM, Tey KE, Chew SF. Clinical profile, morbidity, and outcome of adult-onset generalized pustular psoriasis: analysis of 102 cases seen in a tertiary hospital in Johor, Malaysia. Int J Dermatol 2014;53:676–84.

Choon SE, Navarini AA, Pinter A. Clinical course and characteristics of generalized pustular psoriasis. Am J Clin Dermatol 2022;23(Suppl. 1):21–9.

Choon SE, van de Kerkhof P, Gudjonsson JE, de la Cruz C, Barker J, Morita A, et al. International consensus definition and diagnostic criteria for generalized pustular psoriasis from the International Psoriasis Council. JAMA Dermatol 2024;160:758–68.

Cicchetti DV. [guidelines]. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284–90.

Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. Columbus, OH: Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2014; 2014.

DANDerm [home page]; 2024, https://danderm.dk/ (accessed 25 September 2024).

DermNET [home page]; 2024, https://dermnetnz.org/ (accessed 25 September 2024).

Dermatology Atlas [home page]; 2024, https://atlasdermatologico.com.br/ (accessed 25 September 2024).

DermIS [home page]; 2024, https://www.dermis.net/dermisroot/en/home/index.htm (accessed 25 September 2024).

DermQuest [home page], http://dermquest.com/ (unavailable as of 2026/02/09).

England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. AJR, Am J Roentgenol 2019;212:513–9.

Escalé-Besa A, Vidal-Alaball J, Miró Catalina Q, Gracia VHG, Marin-Gomez FX, Fuster-Casanovas A. The use of artificial intelligence for skin disease diagnosis in primary care settings: a systematic review. Healthcare (Basel) 2024;12:1192.

Fujita H, Terui T, Hayama K, Akiyama M, Ikeda S, Mabuchi T, et al. Japanese guidelines for the management and treatment of generalized pustular psoriasis: the new pathogenesis and treatment of GPP. J Dermatol 2018;45:1235–70.

Gooderham MJ, Van Voorhees AS, Lebwohl MG. An update on generalized pustular psoriasis. Expert Rev Clin Immunol 2019;15:907–19.

Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Genet Program Evolvable Mach 2018;19:305–7.

Havelin A, Hampton P. Telemedicine and e-health in the management of psoriasis: improving patient outcomes - a narrative review. Psoriasis (Auckl) 2022;12:15–24.

Hernández Montilla I, Medela A, Mac Carthy T, Aguilar A, Gómez Tejerina P, Vilas Sueiro A, et al. Automatic International Hidradenitis suppurativa Severity Score System (AIHS4): a novel tool to assess the severity of hidradenitis suppurativa using artificial intelligence. Skin Res Technol 2023;29:e13357.

Huang K, Wu X, Li Y, Lv C, Yan Y, Wu Z, et al. Artificial intelligence-based psoriasis severity assessment: real-world study and application. J Med Internet Res 2023;25:e44932.

HuggingFace.com. Model card for caformer_s18.sail_in1k, https://huggingface.co/timm/caformer_s18.sail_in1k; 2024a (accessed 25 September 2024).

HuggingFace.com. Model card for convformer_s18, sail_in1k, https://huggingface.co/timm/convformer_s18.sail_in1k; 2024b (accessed 25 September 2024).

HuggingFace.com. Model card for convnextv2_nano.fcmae_ft_in1k, https://huggingface.co/timm/convnextv2_nano.fcmae_ft_in1k; 2024c (accessed 25 September 2024).

HuggingFace.com. Model card for convnextv2_tiny.fcmae_ft_in22k_in1k, https://huggingface.co/timm/convnextv2_tiny.fcmae_ft_in22k_in1k; 2024d (accessed 25 September 2024).

HuggingFace.com. Model card for mobilenetv4_conv_large.e500_r256_in1k, https://huggingface.co/timm/mobilenetv4_conv_large.e500_r256_in1k; 2024e (accessed 25 September 2024).

HuggingFace.com. Model card for mobilenetv4_conv_medium.e500_r256_in1k, https://huggingface.co/timm/mobilenetv4_conv_medium.e500_r256_in1k; 2024f (accessed 25 September 2024).

HuggingFace.com. Model card for mobilenetv4_hybrid_large.e600_r384_in1k, https://huggingface.co/timm/mobilenetv4_hybrid_large.e600_r384_in1k; 2024g (accessed 25 September 2024).

HuggingFace.com. Model card for mobilenetv4_hybrid_medium, e500_r224_in1k, https://huggingface.co/timm/mobilenetv4_hybrid_medium.e500_r224_in1k; 2024h (accessed 25 September 2024).

HuggingFace.com. Model card for resnet18.a1_in1k, https://huggingface.co/timm/resnet18.a1_in1k; 2024i (accessed 25 September 2024).

HuggingFace.com. Model card for resnet34.a1_in1k, https://huggingface.co/timm/resnet34.a1_in1k; 2024j (accessed 25 September 2024).

HuggingFace.com. Model card for tf_efficientnetv2_m.in21k_ft_in1k, https://huggingface.co/timm/tf_efficientnetv2_m.in21k_ft_in1k; 2024k (accessed 25 September 2024).

HuggingFace.com. Model card for tf_efficientnetv2_s.in1k, https://huggingface.co/timm/tf_efficientnetv2_s.in1k; 2024l (accessed 25 September 2024).

Huynh QT, Nguyen PH, Le HX, Ngo LT, Trinh NT, Tran MTT, et al. Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. Diagnostics (Basel) 2022;12:1879.

Kharawala S, Golembesky AK, Bohn RL, Esser D. The clinical, humanistic, and economic burden of generalized pustular psoriasis: a structured review. Expert Rev Clin Immunol 2020;16:239–52.

Kovarik C. Development of high-quality artificial intelligence in dermatology: guidelines, pitfalls, and potential [guidelines]. JID Innov 2022;2:100157.

Legit. Health [home page]; 2024, https://legit.health/ (accessed 30 August 2024).

Liu S, Liang Y, Gitter A. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33; 2019. p. 9977–8.

Mac Carthy T, Hernández Montilla I, Aguilar A, García Castro R, González Pérez AM, Vilas Sueiro A, et al. Automatic urticaria activity score: deep learning-based automatic hive counting for urticaria severity assessment. JID Innov 2024;4:100218.

Martorell A, Martin-Gorgojo A, Ríos-Viñuela E, Rueda-Carnero JM, Alfageme F, Taberner R. Artificial intelligence in dermatology: a threat or an opportunity? Actas Dermosifiliogr 2022;113:30–46.

Medela A, Mac Carthy T, Aguilar Robles SA, Chiesa-Estomba CM, Grimalt R. Automatic SCOring of Atopic Dermatitis using deep learning: a pilot study. JID Innov 2022;2:100107.

Morita A, Strober B, Burden AD, Choon SE, Anadkat MJ, Marrakchi S, et al. Efficacy and safety of subcutaneous spesolimab for the prevention of generalised pustular psoriasis flares (Effisayil 2): an international, multicentre, randomised, placebo-controlled trial. Lancet 2023;402:1541–51.

Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array 2022;16:100258.

Navarini AA, Burden AD, Capon F, Mrowietz U, Puig L, Köks S, et al. European consensus statement on phenotypes of pustular psoriasis. J Eur Acad Dermatol Venereol 2017;31:1792–9.

Pan K, Hurault G, Arulkumaran K, Williams HC, Tanaka RJ. EczemaNet: automating detection and severity assessment of atopic dermatitis. In: Liu M, Yan P, Lian C, Cao X, editors. Machine learning in medical imaging. Lecture Notes in Computer Science, 12436. Cham, Switzerland: Springer; 2020. p. 220–30.

Polat G, Ergenc I, Kani HT, Alahdab YO, Atug O, Temizel A. Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. Annual Conference on Medical Image Understanding and Analysis. Cham: Springer International Publishing; 2022. p. 157–71.

Prinz JC, Choon SE, Griffiths CEM, Merola JF, Morita A, Ashcroft DM, et al. Prevalence, comorbidities and mortality of generalized pustular psoriasis: a literature review. J Eur Acad Dermatol Venereol 2023;37:256–73.

Puig L, Choon SE, Gottlieb AB, Marrakchi S, Prinz JC, Romiti R, et al. Generalized pustular psoriasis: a global Delphi consensus on clinical course, diagnosis, treatment goals and disease management. J Eur Acad Dermatol Venereol 2023;37:737–52.

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, et al. DSM-5 field trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. Am J Psychiatry 2013;170:59–70.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618–26.

Tang H, Liu J, Zhao M, Gong X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. Proceedings of the 14th ACM conference on recommender systems. 2020. p. 269–78.

Wang J, Luo Y, Wang Z, Hounye AH, Cao C, Hou M, et al. A cell phone app for facial acne severity assessment. Appl Intell (Dordr) 2023;53:7614–33.

Wiala A, Ranjan R, Schnidar H, Rappersberger K, Posch C. Automated classification of hidradenitis suppurativa disease severity by convolutional neural network analyses using calibrated clinical images. J Eur Acad Dermatol Venereol 2024;38:576–82.

Yu W, Si C, Zhou P, Luo M, Zhou Y, Feng J, et al. MetaFormer baselines for vision. IEEE Trans Pattern Anal Mach Intell 2023;46:896–912.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/