JEADV Clinical Practice

JEADV CLINICAL PRACTICE OPEN ACCESS JOURNAL OF THE EUROPEAN ACADEMY OF DERMATOLOGY & VENEREOLOGY

ORIGINAL ARTICLE OPEN ACCESS

Artificial Intelligence-Based Quantification to Assess the Automatic Psoriasis Area and Severity Index

Taig Mac Carthy[1] | Daniel Dagnino[2] | Alfonso Medela[2] | Gerardo Fernández[1] | Andy Aguilar[1] | Antonio Martorell[1,3] | Pedro Gómez-Tejerina[1] | Gastón Roustán-Gullón[4]

[1]Department of Clinical Endpoint Innovation, Legit.Health, Bilbao, Spain | [2]Department of Medical Data Science, Legit.Health, Bilbao, Spain | [3]Department of Dermatology, Hospital de Manises, Manises, Spain | [4]Dermatology Unit, Puerta de Hierro Hospital, Madrid, Spain

Correspondence: Alfonso Medela (alfonso@legit.health)

Received: 11 March 2025 | **Revised:** 9 June 2025 | **Accepted:** 31 July 2025

Funding: The study was supported by the European Union and the "Centro para el Desarrollo Tecnológico y la Innovación" (CDTI) (SoE-20221020, NextGenEU).

Keywords: deep learning | psoriasis | Psoriasis Area and Severity Index | severity assessment

ABSTRACT

Background: Psoriasis is a chronic inflammatory skin condition affecting millions globally, traditionally assessed via the Psoriasis Area and Severity Index (PASI). Despite its widespread use, PASI suffers from subjectivity, limited sensitivity for small lesions, and high interobserver variability. Recent advances in artificial intelligence offer a promising route to overcome these limitations.

Objectives: This study aimed to develop and validate the Automatic Psoriasis Area and Severity Index (APASI), an AI-driven system for rapid, objective, and standardised assessment of psoriasis severity.

Methods: We compiled the Legit.Health-PsO-PASI dataset, consisting of 2857 images spanning various disease severities and annotated by three expert dermatologists. Two deep learning modules were developed: one for lesion segmentation and another for visual sign intensity classification (erythema, induration and desquamation). Multiple encoder architectures, including ResNet, SE_ResNeXt, Xception, Inceptionv4, EfficientNet, and SegFormer, were evaluated using a five-fold cross-validation approach. Performance metrics included accuracy, root mean square error, intersection over union, confusion matrices and Cohen's kappa for assessing interobserver variability.

Results: The best-performing model, MiT_b2, achieved human-comparable performance, with accuracies of 60.6% for erythema, 54.3% for induration, and 61.8% for desquamation. For lesion segmentation, the Xception model outperformed expert dermatologists, achieving an intersection over union of 0.752.

Conclusions: APASI provides a robust AI-driven framework for psoriasis severity assessment, delivering rapid, objective, and precise evaluations. Its integration into clinical and research workflows could enhance disease monitoring, improve treatment assessment and reduce evaluation costs.

1 | Introduction

Psoriasis, a chronic proliferative and inflammatory skin condition, affects approximately 125 million individuals worldwide [1], with prevalence rates ranging from 0.14% to 1.99% across ethnicities [2, 3]. Characterised by erythematous plaques with silvery scales, psoriasis predominantly manifests on extensor surfaces, the scalp, and the lumbosacral region [4]. The disease

Abbreviations: APASI, Automatic Psoriasis Area and Severity Index; BSA, body surface area; CAD, Computer-aided diagnosis; CM, confusion matrix; FPN, feature pyramid network; GT, ground truth; IoU, intersection over union; PASI, Psoriasis Area and Severity Index; RMSE, root mean square error.

significantly impairs patients' quality of life, causing physical discomfort, psychological distress and functional limitations [5].

Despite the clinical visibility of psoriasis lesions, objective and reliable measures to ensure consistent severity assessments remain critical. This challenge is particularly pronounced in clinical trials, where interobserver variability in interpreting visual signs compromises evaluations [6]. The Psoriasis Area and Severity Index (PASI) [7], the gold standard for severity assessment, suffers from limitations such as low sensitivity to small affected areas and significant interobserver variability [8]. Simpler tools like the Physician Global Assessment (PGA) enable rapid evaluations but lack granularity and depend heavily on subjective expertise [9]. While alternative systems, such as the lattice system [10] and the National Psoriasis Foundation Psoriasis Score [11, 12], have sought to improve objectivity, none have supplanted PASI in clinical practice. Recent advances in artificial intelligence (AI) demonstrate promise in reducing interobserver variability and enabling automation [13–15].

In recent years, deep learning algorithms have been applied to psoriasis, focusing primarily on lesion detection or non-standardised severity score assessment. For example, Shrivastava et al. [16] achieved high accuracy in classifying localised psoriasis images into five severity categories using support vector machines (SVM), neural networks, and decision tree classifiers combined with feature selection. Munia et al. [17] developed an automated approach for lesion segmentation and plaque localisation using a SVM with a novel colour-based feature extraction method. Similarly, George et al. [18] employed super-pixel decomposition followed by k-means clustering to segment lesions, effectively distinguishing between lesion and non-lesion skin. Dash et al. [19] demonstrated that convolutional neural networks based on a modified U-Net architecture could segment psoriasis lesions with high accuracy, sensitivity, and specificity compared to traditional approaches.

While some works have extended beyond segmentation to include severity assessment, many do not rely on standardised scoring methods, limiting their clinical applicability. For instance, Dash et al. [20] proposed a computer-aided diagnosis (CAD) system that integrates segmentation and classification into four severity levels; however, the classification framework was not grounded in a widely accepted clinical standard such as the Psoriasis Area and Severity Index (PASI). Notably, only Huang et al. [21] incorporated a standardised method for PASI scoring, offering an important step toward clinical integration. However, their approach did not include a detailed evaluation of specific visual signs, which are critical for understanding the practical benefits and limitations of automated systems compared to human expert assessments. This underscores the need for more comprehensive AI-based solutions that align closely with clinical standards and facilitate transparent, interpretable comparisons with healthcare practitioners.

We present the Automatic Psoriasis Area and Severity Index (APASI), a fully automated adaptation of PASI designed for clinical and research use. APASI delivers rapid, objective, and reproducible assessments of psoriasis severity from clinical images. Its intended use is to support healthcare providers during the evaluation of psoriatic lesions by automatically quantifying disease severity, thereby enhancing the efficiency, accuracy, and standardisation of care delivery within the dermatological clinical workflow.

The APASI tool presented here is accessible through the Legit.Health platform. Additionally, to promote transparency and academic collaboration, we provide a detailed technical description of the methodology and the dataset used, enabling reproducibility and further research by the scientific community.

2 | Materials and Methods

2.1 | APASI

The PASI measures the average erythema, induration, and desquamation of the lesions weighted by the area of involvement. The PASI score can be calculated as follows:

$$PASI = 0.1(E_h + I_h + D_h)A_h + 0.3(E_t + I_t + D_t)A_t + 0.2(E_u + I_u + D_u)A_u + 0.4(E_l + I_l + D_l)A_l$$

where E = erythema, I = induration, D = desquamation and A = area. The sites are h, head; t, trunk; u, upper extremities; and l, lower extremities. A numerical value is given to the extent of the lesions in these areas; 0, 0%; 1, 1%–9%; 2, 10%–29%; 3, 30%–49%; 4, 50%–69%; 5, 70%–89%; 6, 90%–100%. E, I, and D are assessed on a 5-point scale: 0, no symptoms; 1, mild; 2, moderate; 3, severe; and 4, very severe.

To automate PASI, we developed the APASI, a deep learning model that analyzes single smartphone/tablet images to measure the extent of psoriasis and evaluate the intensity of erythema, induration, and desquamation.

2.2 | Deep Learning-Based Models

APASI comprises two main modules (Legit.Health-APASInet-research): one for lesion surface segmentation and another for visual sign intensity classification. Two architectural variants were developed:

1. A shared Feature Pyramid Network (FPN) encoder [22] architecture, where extracted features are used for: (a) classifying visual sign intensities, and (b) lesion segmentation at a resolution of 512×512px
2. An architecture with two independent FPN encoders for each task, and the same decoders as in the previous model (1).

And various FPN encoder architectures namely ResNet [23], SE_ResNeXt [24, 25], Xception [26], Inception [27], EfficientNet [28] and Mix Vision Transformer (MiT) based on SegFormer [29] were evaluated.

2.3 | Datasets and Annotations

This retrospective, non-interventional study presents the Legit.Health-PsO-PASI dataset, developed specifically to train

and validate APASI. The dataset consists of 2857 images of psoriasis, sourced from public domains and automatically filtered by diagnosis, using metadata provided in the original atlases. These images represent a broad range of clinical scenarios, affected body regions, and disease severities, primarily involving patients with Fitzpatrick skin types I–IV (92%), with 8% corresponding to types V and VI. Additionally, for model training, we incorporated 1364 publicly available images of healthy skin from the internet.

The data set was organised into two subgroups:

- Dataset-S (1364 images): In this subgroup, three independent annotators provided detailed annotations for both lesion segmentation and visual sign intensities. These annotators were non-healthcare professionals who received a structured annotation guide with clear visual examples and definitions for each PASI-related visual sign. Notably, the analysis revealed a limited number of samples at the highest intensity level (degree 4): 20 for erythema, 19 for desquamation, and only 6 for induration.
- Dataset-VS (2500 images): To address the underrepresentation of higher severity cases and achieve a more balanced distribution of severity levels, this subgroup was created and annotated exclusively for visual sign intensities by four independent expert dermatologists, one with 7 years of experience and three with over 10 years of experience.

Ground truth (GT) labels for both lesion segmentation and visual sign intensity classification were generated by averaging the annotations.

2.4 | Evaluation Metrics

To quantify interobserver variability, we used Cohen's kappa to assess pairwise agreement for erythema, induration, and desquamation grades. For segmentation, we evaluated agreement using the Intersection over Union (IoU) metric.

To benchmark model performance, we measured accuracy, and root mean square error (RMSE), and analysed confusion matrices (CM) for visual sign grades and IoU for segmentations.

2.5 | Experimental Setup

A five-fold cross-validation strategy was applied to train the models, ensuring robust performance evaluation across different folds.

Dataset-S was extended with images of healthy skin (named Dataset-S-EHS). These healthy images were used exclusively for model training to improve the model's ability to accurately detect healthy skin while ensuring that the evaluation remained unbiased by excluding these images from the validation and test sets.

In Dataset-VS, we observed a strong right-skewed distribution, particularly with a peak at score 2 across erythema and induration. To balance the sampling of severity levels across all classes, we extended Dataset-VS, which originally contained 62 samples with a severity score of 0 for erythema. We created the Dataset-VS-EHS, by adding 20% of healthy images, thereby increasing the representation of low-severity cases (Supporting Information: Figure S1).

All models used in this study were initialised with weights pretrained on the ImageNet dataset [30]. The training configuration was standardised across all models, with each trained for 100 epochs using a batch size of 32 and the Adam optimiser. A constant learning rate decay schedule was implemented, gradually reducing the learning rate from 10^{-3} to 10^{-4} to ensure stable convergence and optimal performance. The training process employed a Cross-Entropy loss function.

2.6 | CAD System

To make the algorithms accessible to healthcare professionals, we created a fully integrated CAD system, a web application that integrates the Legit.Health-APASInet-production algorithm to automatically calculate the patient-specific APASI.

In the first stage, images of affected areas are uploaded to the system. The user has to choose the body zone from the options defined in the original PASI: head, trunk, upper limbs, and lower limbs. In the second stage, the algorithm calculates the severity of psoriasis. Finally, the algorithm output is shown in a user-friendly report containing an image with the estimated lesion surface and a chart showing the evolution of the APASI over time (Figure 1).

Computing the APASI requires calculating the proportion of skin covered by the lesion in the body zones defined in the PASI. We solved this by including a small piece of hardware called AI Marker, a sticker with several shapes and colours that help translate pixels into a metric unit. In addition, the Body Surface Area (BSA) is calculated using the patient's height and weight using the Mosteller [31, 32] formula. Once the surface of the lesion and BSA are estimated, the percentage can be calculated by dividing the lesion surface by the BSA of the body zone.

The final formula for N images of the whole body can be written as follows:

$$APASI = \sum_{i=1}^{N} A_i W_{zone} \Big(\max(E_1, ..., E_{N_{zone}}) + \max(I_1, ..., I_{N_{zone}}) + \max(D_1, ..., D_{N_{zone}}) \Big)$$

where

$$S_i = \frac{a_i}{BSAW_{zone}}$$

and

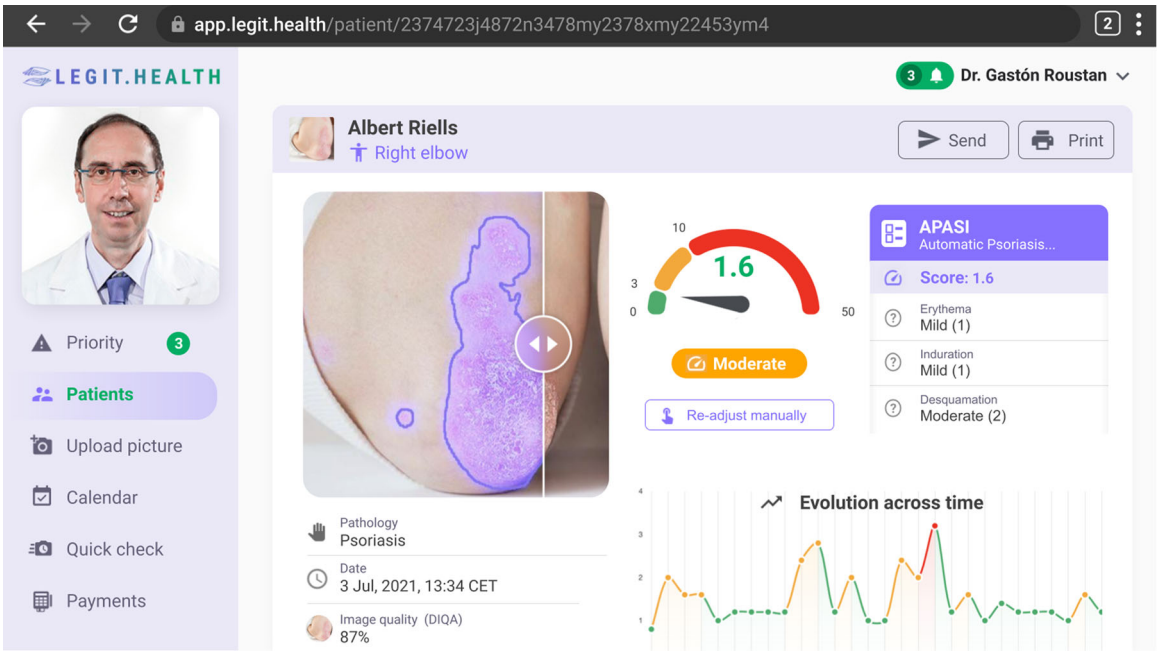

FIGURE 1 | Example report generated by the CAD system. The interface allows doctors to assess the lesion surface by moving the slider to toggle between the original image and the image highlighting the estimated lesion area. The chart below shows the evolution of the APASI over time.

$$A_i = \begin{Bmatrix} 0 if S_i = 0 \\ 1 if S_i = [1,10] \\ 2 if S_i = [11,29] \\ 3 if S_i = [31,50] \\ 4 if S_i = [51,70] \\ 5 if S_i = [71,90] \\ 6 if S_i = [91,100] \end{Bmatrix}$$

N is the total number of images, A_i [0, 6] stands for the lesion area categorical score, S_i [0, 100] stands for surface percentage or lesion surface divided by body surface area of the body zone, N_{zone} is the number of images of a body zone, $W_{\text{zone}} = 0.1$, 0.2, 0.3, 0.4 is the weight of the body zone of i-th lesion and E_k, I_k and D_k are erythema, induration, and desquamation intensity values in the range [0–4].

3 | Results

3.1 | Annotations' Variability Analysis

Cohen's kappa coefficient on Dataset-VS revealed relatively weak agreement among annotators, with kappa values of 0.265 for erythema, 0.274 for induration, and 0.384 for desquamation. However, when accounting for the ordinal nature of the annotated intensity values using quadratic weighting, the kappa coefficients increased significantly to 0.643 for erythema, 0.684 for induration, and 0.758 for desquamation.

For Dataset-S, a similar analysis showed low exact agreement among annotators, with kappa values of 0.149 for erythema, 0.067 for induration, and 0.104 for desquamation. After applying quadratic weighting, the kappa coefficients improved to 0.442 for erythema, 0.266 for induration, and 0.375 for desquamation. While this adjustment enhanced agreement, the results still indicate relatively low concordance among annotators.

Regarding lesion segmentation, when using masks defined by all three annotators to establish the GT, the IoU achieved a value of 0.693 in a one-versus-rest approach.

Based on this variability analysis, Dataset-VS was used to train models for classifying the severity of visual signs, while Dataset-S was employed for segmentation tasks.

3.2 | Model Evaluation

3.2.1 | Visual Signs Intensity Classification

3.2.1.1 | Accuracy and RMSE Results. Table 1 presents the performance of models trained on Dataset-VS, revealing a consistent trend: models achieved the highest performance on erythema classification, while induration exhibited the lowest performance, with desquamation falling in between. This disparity is likely due to the nature of induration, which involves an inflammatory response with partial three-dimensional characteristics, making it more challenging to assess from two-dimensional images.

Among the evaluated architectures, the MiT-based models, specifically, MiT_b3-clf and MiT_b2-clf, emerged as top performers. For erythema classification, MiT_b3-clf achieved accuracy that surpassed that of human annotators. In the case of induration, although the model attained higher accuracy, it also exhibited a higher RMSE compared to the annotators, suggesting greater prediction variability. Conversely, for desquamation, both accuracy and RMSE metrics were inferior to those of the human annotators.

TABLE 1 | Visual sign intensity classification results on models trained on Dataset-VS.

	Erythema		Induration		Desquamation	
Encoder or annotator	**Accuracy**	**RMSE**	**Accuracy**	**RMSE**	**Accuracy**	**RMSE**
Annotators (one-vs-rest)	0.525	0.795	0.516	0.820	0.590	0.762
EfficientNet-b2-clf	0.59	0.769	0.469	0.939	0.501	1.01
EfficientNet-b3-clf	0.563	0.814	0.461	0.927	0.533	0.986
Inceptionv4-clf	0.574	0.8	0.502	0.888	0.518	0.987
MiT_b1-clf	0.612	0.716	0.512	0.86	0.56	0.887
MiT_b2-clf	0.61	0.715	0.53	0.828	0.574	0.853
MiT_b3-clf	0.614	0.694	0.522	0.872	0.584	0.87
ResNet18-clf	0.605	0.746	0.457	0.939	0.509	1.00
ResNet34-clf	0.603	0.76	0.478	0.933	0.51	1.01
SE_ResNeXt101_32x4d-clf	0.63	0.714	0.509	0.88	0.526	0.922
SE_ResNeXt50_32x4d-clf	0.59	0.757	0.499	0.878	0.538	0.94
Xception-clf	0.568	0.837	0.446	0.926	0.478	1.05

Note: Performance comparison of various encoder architectures and human annotators for erythema, induration, and desquamation classification. The annotators (one-vs-rest) serve as a benchmark for interobserver agreement.
Abbreviation: RMSE, root mean square error.

TABLE 2 | Visual sign intensity classification results on models trained on Dataset-VS-EHS.

	Erythema		Induration		Desquamation	
Encoder or annotator	**Accuracy**	**RMSE**	**Accuracy**	**RMSE**	**Accuracy**	**RMSE**
Annotators (one-vs-rest)	0.525	0.795	0.516	0.820	0.590	0.762
EfficientNet-b2-clf	0.588	0.788	0.477	0.963	0.594	0.909
EfficientNet-b3-clf	0.551	0.88	0.509	0.947	0.599	0.874
Inceptionv4-clf	0.574	0.825	0.504	0.934	0.594	0.862
MiT_b1-clf	0.591	0.76	0.515	0.868	0.615	0.816
MiT_b2-clf	0.606	0.748	0.543	0.856	0.618	0.81
MiT_b3-clf	0.6	0.761	0.507	0.908	0.609	0.812
ResNet18-clf	0.558	0.915	0.472	1.07	0.573	0.921
ResNet34-clf	0.555	0.924	0.452	1.03	0.571	0.894
SE_ResNeXt101_32x4d-clf	0.567	0.841	0.504	0.946	0.601	0.856
SE_ResNeXt50_32x4d-clf	0.586	0.837	0.479	0.967	0.579	0.896
Xception-clf	0.567	0.843	0.442	0.978	0.57	0.898

Note: Performance comparison of various encoder architectures and human annotators for erythema, induration, and desquamation classification. The annotators (one-vs-rest) serve as a benchmark for interobserver agreement.
Abbreviation: RMSE, root mean square error.

Other architectures, such as EfficientNet, Xception, and Inceptionv4, exhibited lower overall performance compared to ResNet and its variant SE_ResNeXt.

Table 2, which evaluates models trained on Dataset-VS-EHS, shows a similar trend: induration remained the most challenging category, but the performance gap between erythema and desquamation was narrower. In this evaluation, MiT_b2-clf demonstrated the best overall performance across all metrics and visual signs, outperforming annotators in terms of accuracy.

3.2.1.2 | Confusion Matrix Analysis. To further investigate the differences between annotators' assessments and model predictions, Figure 2 presents the confusion matrices for visual sign intensity classification. The results correspond to the MiT_b3-clf model trained on Dataset-VS and the MiT_b2-clf model trained on Dataset-VS-EHS, which achieved the best performance on their respective datasets.

3.2.1.3 | Analysis of the Annotators' Confusion Matrix Results.

- Erythema (Figure 2a): The lowest agreement among annotators was observed at severity 0, with only 35% accuracy. Notably, a single annotator often predicts one level higher than the GT 40% of the time, and even 3 levels higher 15% of the time. The highest agreement was for severity 4, at

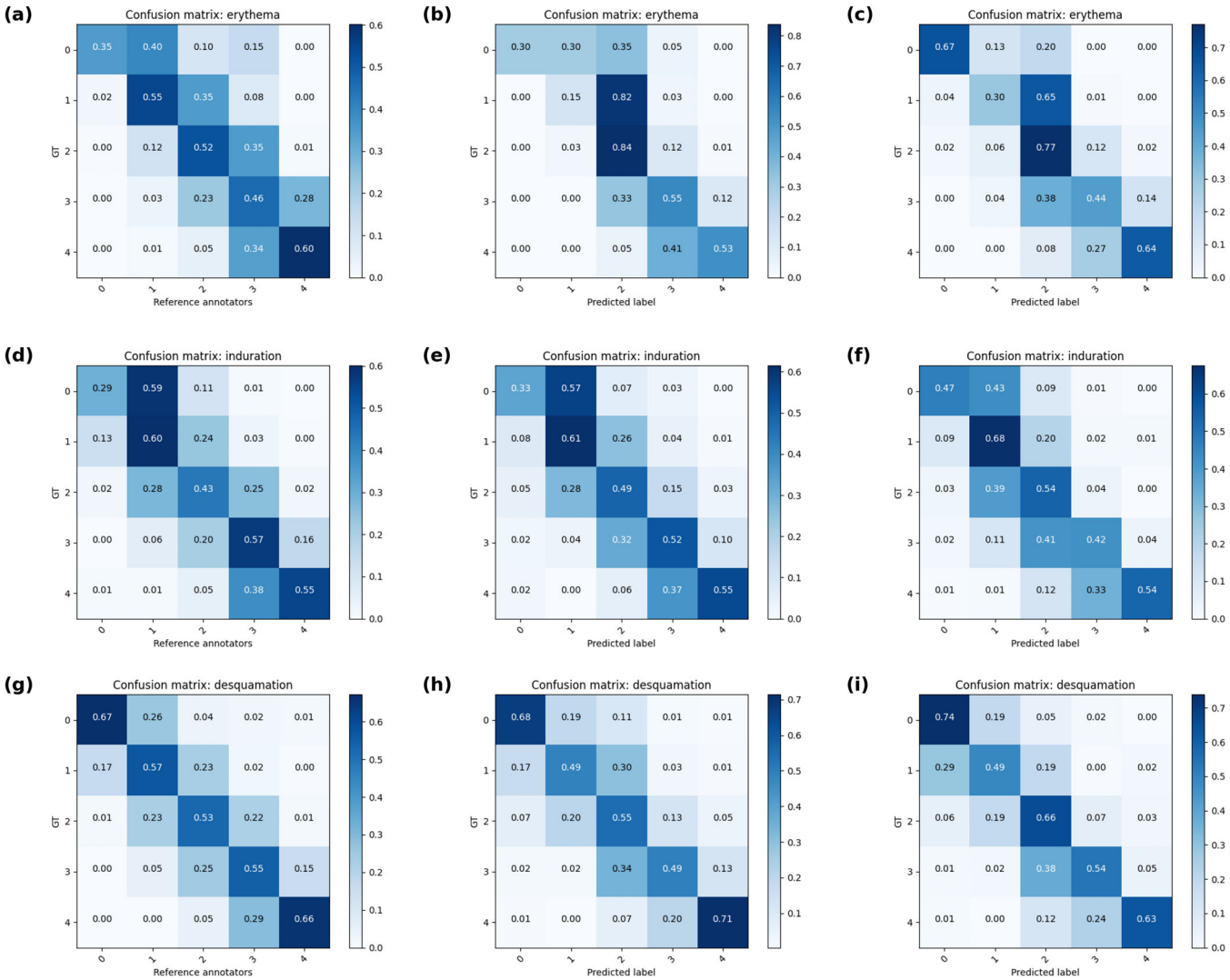

FIGURE 2 | Confusion matrices for visual sign intensity classification. The left column (a, d, g) represents the mean CMs of the annotators, computed using a one-versus-rest approach. The middle column (b, e, h) displays results from one of the best-performing models, MiT-b3, trained on Dataset-VS. The right column (c, f, i) shows results from another top-performing model, MiT-B2, trained on Dataset-VS-EHS. Each row corresponds to a different visual sign: erythema (a–c), induration (d–f), and desquamation (g–i).

60%. When considering a ±1 severity range (tri-diagonal), the match was over 90% for all severities except 0, where it was 75%. This strengthens the adequateness of Cohen's Kappa as a more suitable metric for evaluating annotator performance on this ordinal data.

- Induration (Figure 2d): The lowest agreement was at severity 0, with only 29% accuracy, and 59% of cases predicted one severity higher than the GT. In 11% of cases, predictions were two levels higher. For severities above 0, the match within ±1 severity exceeded 93%.
- Desquamation (Figure 2g): The annotators showed a strong match, with all diagonal scores above 0.53, and the ±1 severity match exceeding 93%.

3.2.1.4 | Analysis of the Model's Confusion Matrix. For the MiT_b3-clf model trained on Dataset-VS (Figure 2b,e,h), there was a clear asymmetry in its CM. For example, erythema severity 1 was often confused with severity 2 (82% of the time), and severity 0 shows a score of just 30%. This reflects the skewness in Dataset-VS However, when trained MiT_b2-clf on Dataset-VS-EHS (Figure 2c,f,i), this skewness was largely mitigated. The match at severity 0 improved to 67%, and the ±1 severity match raised to 80% at severity 0 and exceeded 92% for higher severities.

For induration, the MiT_b2-clf model showed diagonal scores above 0.42, and the ±1 severity match was 88%, indicating a strong match between model predictions and GT. Similarly, for desquamation, the diagonal scores exceeded 0.49, with the ±1 severity match surpassing 80%.

These results demonstrated that model predictions, particularly those from MiT_b2-clf, performed on par with human annotators, in some cases surpassing them. Moreover, training models with reduced severity skewness significantly improved model performance. Notably, for severe cases, such as those with severity 4, human annotators exhibited a relatively low rate of misclassification as severity 2, specifically 5% for erythema, induration, and desquamation. In contrast, the MiT_b2-clf model showed higher misclassification rates in these cases,

reaching 8%, 12%, and 12%, respectively, indicating that the model underperforms compared to human annotators in these specific cases.

3.3 | Lesion Segmentation

Table 3 presents the evaluation metrics, specifically IoU and F1-score, for all trained models. The models are identified by their encoder architecture, with the suffixes "-seg" or "-multi" indicating whether the encoder was trained solely for the segmentation task ("-seg") or in a multi-task fashion for both lesion segmentation and the classification of visual sign intensity ("-multi"). All models were trained on the Dataset-S-EHS. The use of visual sign annotations in the "-multi" models allowed us to assess the impact of multi-task learning on final performance metrics. In addition to model performance metrics, Table 3 includes results obtained from annotators for the same metrics and test dataset.

This benchmark revealed that the model employing an Xception encoder within a multi-task core architecture referred to as Xception-multi, achieved the highest performance. Specifically, Xception-multi attained an IoU of 0.752 and an F1-score of 0.843 in the segmentation task. Notably, its IoU surpassed that of the annotators (0.702), reinforcing the effectiveness of multi-task learning, while its F1-score was only 6% below that of expert dermatologists.

Additionally, models trained with a multi-task loss function consistently outperformed those trained with a segmentation-only loss. We attribute this improvement to the inclusion of lesion-free images, where the classification task helps mitigate false positives in segmentation by reducing pixel-level errors.

TABLE 3 | Lesion segmentation performance comparison.

Method	IoU	F1-score
Annotators (one-vs-rest)	0.702	0.896
Xception-multi	0.752	0.843
EfficientNet-b2-multi	0.742	0.836
SE_ResNeXt50_32x4d-multi	0.737	0.83
Inceptionv4-multi	0.714	0.813
ResNet18-multi	0.706	0.807
EfficientNet-b3-multi	0.703	0.798
EfficientNet-b2-seg	0.699	0.788
ResNet34-multi	0.692	0.793
EfficientNet-b3-seg	0.669	0.763
MiT_b3-seg	0.65	0.757
MiT_b1-seg	0.646	0.755
Xception-seg	0.639	0.745
MiT_b2-seg	0.622	0.735
ResNet34-seg	0.58	0.687
ResNet18-seg	0.562	0.666
inceptionv4-seg	0.546	0.649
MiT_b2-multi	0.517	0.637
SE_ResNeXt50_32x4d-seg	0.483	0.578
MiT_b3-multi	0.463	0.576
MiT_b1-multi	0.218	0.268

Note: Performance comparison of various encoder architectures and human annotators for lesion segmentation. Annotators (one-vs-rest) serve as a benchmark for interobserver agreement.

4 | Discussion

This study demonstrates the potential of the evaluated technology to transform psoriasis severity assessments. Unlike traditional manual PASI scoring and other established systems, APASI enables rapid, objective evaluations and should be further developed as a complementary tool to human assessment. By providing a detailed, automated, and standardised scoring system, APASI stands to benefit patients, clinicians, and the pharmaceutical industry alike. Moreover, because image capture does not require a physician's direct involvement, this approach could reduce dermatology costs by saving time and enhancing the documentation of disease progression. This is particularly valuable in pharmaceutical clinical trials, where improved scoring accuracy may yield more granular and reliable data than what is achievable with classic systems.

To our knowledge, this is the first study to automatically assess psoriasis severity using the PASI scoring system as a direct reference, producing a clinically interpretable score, providing segmentation of the affected areas, and reporting detailed performance metrics for both classification (erythema, desquamation, and induration) and segmentation models. This comprehensive evaluation allows for a transparent comparison of APASI's performance against expert dermatologists. In contrast to prior works that have focused solely on segmentation or proposed alternative, non-standard severity scales, our approach delivers a unified and clinically aligned solution that reflects routine dermatological practice.

Among the state-of-the-art models evaluated, the MiT_b2-clf achieved accuracy and RMSE levels comparable to those of expert dermatologists, notably outperforming the annotators in terms of accuracy. However, for desquamation and induration, the model exhibited a slightly higher RMSE compared to the experts. Additionally, the Xception-multi model excelled in the segmentation task, achieving an IoU of 0.752, surpassing the expert dermatologists' IoU of 0.702, and its F1-score was only 6% lower than that of the experts.

While this study demonstrates the potential of APASI to transform psoriasis severity assessments through automated, objective, and scalable evaluations, some limitations must be acknowledged. First, although APASI does not require a physician's direct involvement during image capture, accurate scoring relies on comprehensive imaging of all affected areas. This constraint represents a key limitation of the current analysis. To address this, we are planning a clinical evaluation of body area calculation that includes a standardised imaging protocol. Participants will be instructed to capture images of all visibly affected regions, guided by a structured protocol to

ensure comprehensive body coverage. With this approach, it will be possible to approximate the affected body surface area more accurately and derive a more complete APASI score. Future work will focus on validating APASI in real-world clinical workflows.

Another limitation of this study is the underrepresentation of Fitzpatrick skin types V and VI, which comprised only 8% of the dataset. This imbalance may affect the generalisability and accuracy of APASI in patients with darker skin tones, as visual features of psoriasis can present differently across skin types. Recognising this limitation, we are actively and continuously seeking to expand the dataset with more diverse skin types, particularly V and VI, to improve model performance, reduce bias, and enhance the robustness and equity of the system across all patient populations.

5 | Conclusions

APASI represents a transformative AI-driven tool for psoriasis severity assessment, offering objective, rapid, and precise evaluations. Its integration into clinical and research workflows has the potential to standardise scoring, reduce costs, and enhance data quality in both clinical practice and pharmaceutical research. While these initial results are promising, they are based solely on atlas data. To address this, future work will focus on clinical validation through in-person PASI scoring and real-world implementation. A structured imaging protocol will be employed to approximate body surface area involvement more accurately, thereby enabling comprehensive APASI scoring in diverse patient populations. This step is crucial for ensuring the model's generalisability, reliability, and clinical applicability.

Author Contributions

Conceptualisation: Alfonso Medela, Daniel Dagnino, Taig Mac Carthy, and Gaston Roustan-Gullon. Data Curation: Daniel Dagnino, Alfonso Medela, and Pedro Gomez-Tejerina. Formal Analysis: Daniel Dagnino, Alfonso Medela, and Pedro Gomez-Tejerina. Investigation: Daniel Dagnino, Alfonso Medela, Antonio Martorell, Taig Mac Carthy, Pedro Gomez-Tejerina, and Gaston Roustan-Gullon. Methodology: Daniel Dagnino, Alfonso Medela, Antonio Martorell, Taig Mac Carthy, Pedro Gomez-Tejerina, and Gaston Roustan-Gullon. Project Administration: Alfonso Medela, Andy Aguilar, and Gaston Roustan-Gullon. Resources: Alfonso Medela, Taig Mac Carthy, Gerardo Fernandez, and Andy Aguilar. Software: Daniel Dagnino, Alfonso Medela, and Gerardo Fernandez. Validation: Daniel Dagnino, Alfonso Medela, and Taig Mac Carthy. Visualisation: Daniel Dagnino and Taig Mac Carthy. Writing–Original Draft Preparation: Daniel Dagnino, Alfonso Medela, and Gaston Roustan-Gullon. Writing–Review and Editing: Daniel Dagnino, Alfonso Medela, Antonio Martorell, Taig Mac Carthy, Andy Aguilar, Gerardo Fernandez, Pedro Gomez-Tejerina, and Gaston Roustan-Gullon.

Acknowledgements

The authors acknowledge the financial support provided by the European Union and the "Centro para el Desarrollo Tecnológico y la Innovación" (CDTI) (SoE-20221020, NextGenEU) and Legit.Health, AI Labs Group S.L.

Ethics Statement

This study does not include identifiable patient data or images. All images used in this study were sourced from publicly available dermatologic atlases, ensuring compliance with ethical guidelines. No direct patient interaction or data collection was involved, and therefore, informed consent was not required. This study adheres to the "JEADV Clinical Practice" and ICMJE recommendations for patient privacy and ethical research conduct. Ethical Approval: This study did not involve direct human or animal experimentation. The images used were obtained from publicly available dermatologic atlases. As such, institutional ethics committee approval was not required. The research complies with the ethical standards outlined in the Helsinki Declaration.

Conflicts of Interest

D.D. is an employee of Legit.Health. A.A., A.M., G.F., and T.M.C. are shareholders of AI LABS GROUP SL. The remaining authors declare no conflicts of interest.

Data Availability Statement

Legit.Health-PsO-PASI dataset images used in this study article are publicly available at DermNetNZ (https://www.dermnetnz.org), Danderm (https://danderm-pdv.is.kkh.dk/atlas), DermatoWeb (www.dermatoweb.net), Dermatology Atlas (http://www.atlasdermatologico.com.br), Interactive Dermatology Atlas (http://www.dermatlas.net), DermIS (https://www.dermis.net), Dermquest (https://www.dermquest.com), Hellenic Dermatological Atlas (http://www.hellenicdermatlas.com/en), Full-Spectrum-Dermatology (https://jddonline.com/store/The-Full-Spectrum-of-Dermatology-A-Diverse-and-Inclusive-Atlas-Digital-Edition-p450581852), Dermatology Atlas (https://www.meddean.luc.edu/lumen/meded/medicine/dermatology/melton/atlas.htm). Legit.Health-APASInet-research and Legit.Health-APASInet-production are not publicly available.

References

1. A. W. Armstrong and C. Read, "Pathophysiology, Clinical Presentation, and Treatment of Psoriasis: A Review," *Journal of the American Medical Association* 323, no. 19 (2020): 1945–1960.

2. R. Parisi, I. Iskandar, E. Kontopantelis, et al., "National, Regional, and Worldwide Epidemiology of Psoriasis: Systematic Analysis and Modelling Study," *BMJ* 369 (2020): 1590.

3. A. Brandon, A. Mufti, and R. Gary Sibbald, "Diagnosis and Management of Cutaneous Psoriasis: A Review," *Advances in Skin & Wound Care* 32, no. 2 (2019): 58–69.

4. A. C. Gamret, A. Price, R. M. Fertig, H. Lev-Tov, and A. J. Nichols, "Complementary and Alternative Medicine Therapies for Psoriasis: A Systematic Review," *JAMA Dermatology* 154, no. 11 (2018): 1330–1337.

5. S. R. Rapp, S. R. Feldman, M. L. Exum, A. B. Fleischer, and D. M. Reboussin, "Psoriasis Causes as Much Disability as Other Major Medical Diseases," *Journal of the American Academy of Dermatology* 41, no. 3 (1999): 401–407.

6. K. B. Gordon, K. Reich, J. J. Crowley, et al., "Disease Activity and Treatment Efficacy Using Patient-Level Psoriasis Area and Severity Index Scores From Tildrakizumab Phase 3 Clinical Trials," *Journal of Dermatological Treatment* 33, no. 1 (2022): 219–228.

7. T. Fredriksson and U. Pettersson, "Severe Psoriasis–Oral Therapy With a New Retinoid," *Dermatology* 157, no. 4 (1978): 238–244.

8. C. Fink, C. Alt, L. Uhlmann, C. Klose, A. Enk, and H. A. Haenssle, "Intra-And Interobserver Variability of Image-Based PASI Assessments in 120 Patients Suffering From Plaque-Type Psoriasis," *Journal of the European Academy of Dermatology and Venereology* 32, no. 8 (2018): 1314–1319.

9. J. A. Walsh, H. Jones, L. Mallbris, et al., "The Physician Global Assessment and Body Surface Area Composite Tool Is a Simple Alternative to the Psoriasis Area and Severity Index for Assessment of Psoriasis: Post Hoc Analysis From PRISTINE and PRESTA," *Psoriasis: Targets and Therapy* 8 (2018): 65–74.

10. R. G. Langley and C. N. Ellis, "Evaluating Psoriasis With Psoriasis Area and Severity Index, Psoriasis Global Assessment, and Lattice System Physician's Global Assessment," *Journal of the American Academy of Dermatology* 51, no. 4 (2004): 563–569.

11. A. B. Gottlieb, U. Chaudhari, D. G. Baker, M. Perate, and L. T. Dooley, "The National Psoriasis Foundation Psoriasis Score (NPF-PS) system versus the Psoriasis Area Severity Index (PASI) and Physician's Global Assessment (PGA): A Comparison," *Journal of Drugs in Dermatology* 2, no. 3 (2003): 260–266.

12. C. S. Carlin, K. P. Callis, and G. G. Krueger, "Efficacy of Acitretin and Commercial Tanning Bed Therapy for Psoriasis," *Archives of Dermatology* 139, no. 4 (2003): 436–442.

13. A. Medela, T. Mac Carthy, S. A. Aguilar Robles, C. M. Chiesa-Estomba, and R. Grimalt, "Automatic SCOring of Atopic Dermatitis Using Deep Learning: A Pilot Study," *JID Innovations* 2, no. 3 (2022): 100107.

14. I. Hernández Montilla, A. Medela, T. Mac Carthy, et al., "Automatic International Hidradenitis Suppurativa Severity Score System (AIHS4): A Novel Tool to Assess the Severity of Hidradenitis Suppurativa Using Artificial Intelligence," *Skin Research and Technology* 29, no. 6 (2023): e13357.

15. T. Mac Carthy, I. Hernández Montilla, A. Aguilar, et al., "Automatic Urticaria Activity Score: Deep Learning–Based Automatic Hive Counting for Urticaria Severity Assessment," *JID Innovations* 4, no. 1 (2024): 100218.

16. V. K. Shrivastava, N. D. Londhe, R. S. Sonawane, and J. S. Suri, "A Novel and Robust Bayesian Approach for Segmentation of Psoriasis Lesions and Its Risk Stratification," *Computer Methods and Programs in Biomedicine* 150 (2017): 9–22.

17. T. T. Munia, I. R. i Haque, A. Aymond, et al., "Automatic Clustering-Based Segmentation and Plaque Localization in Psoriasis Digital Images." *IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT) 2017 Nov 6* (IEEE, 2017), 113–116.

18. Y. George, M. Aldeen, and R. Garnavi, "Automatic Psoriasis Lesion Segmentation in Two-Dimensional Skin Images Using Multiscale Superpixel Clustering," *Journal of Medical Imaging* 4, no. 4 (2017): 044004.

19. M. Dash, N. D. Londhe, S. Ghosh, A. Semwal, and R. S. Sonawane, "PsLSNet: Automated Psoriasis Skin Lesion Segmentation Using Modified U-Net-Based Fully Convolutional Network," *Biomedical Signal Processing and Control* 52 (2019): 226–237.

20. M. Dash, N. D. Londhe, S. Ghosh, R. Raj, and R. S. Sonawane, "A Cascaded Deep Convolution Neural Network Based CADx System for Psoriasis Lesion Segmentation and Severity Assessment," *Applied Soft Computing* 91 (2020): 106240.

21. K. Huang, X. Wu, Y. Li, et al., "Artificial Intelligence–Based Psoriasis Severity Assessment: Real-World Study and Application," *Journal of Medical Internet Research* 25 (2023): e44932.

22. S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature Pyramid Network For Multi-Class Land Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2018), 272–275.

23. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 770–778.

24. A. G. Roy, N. Navab, and C. Wachinger, "Concurrent Spatial And Channel 'Squeeze & Excitation' in Fully Convolutional Networks." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I* (Springer International Publishing, 2018), 421–429.

25. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 1492–1500.

26. F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition* (IEEE, 2017), 1251–1258.

27. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, et al., "Going Deeper With Convolutions," in *Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition* (IEEE, 2015), 1–9.

28. M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling For Convolutional Neural Networks." *International Conference on Machine Learning* (PMLR, 2019), 6105–6114.

29. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation With Transformers," *Advances in Neural Information Processing Systems* 34 (2021): 12077–12090.

30. O. Russakovsky, J. Deng, H. Su, et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115 (2015): 211–252.

31. A. Orimadegun and A. Omisanjo, "Evaluation of Five Formulae for Estimating Body Surface Area of Nigerian Children," *Annals of Medical and Health Sciences Research* 4, no. 6 (2014): 889–898.

32. J. Y. Lee, J. W. Choi, and H. Kim, "Determination of Body Surface Area and Formulas to Estimate Body Surface Area Using the Alginate Method," *Journal of Physiological Anthropology* 27, no. 2 (2008): 71–82.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Figure S1: **Frequency distribution of desquamation, erythema, and induration scores across the training and validation datasets, and the test dataset for both Dataset-VS and Dataset-VS-EHS.** The blue bars represent the frequency of scores in the combined training and validation split of Dataset-VS. The orange bars represent the same for Dataset-VS-EHS. The green bars show the frequency of scores in the test dataset, which remains consistent across both Dataset-VS and Dataset-VS-EHS. Each panel focuses on a specific score type: Desquamation (a), Erythema (b), and Induration (c). The x-axis shows the severity levels, while the y-axis shows the frequency of occurrences.