Skin Health and Disease, 2026, 00, 1–8
https://doi.org/10.1093/skinhd/vzag062
Published: 9 June 2026
Original article

SHD OXFORD
Skin Health and Disease

ALADIN (Acne Lesion And Density INdex): a novel tool for automatic acne severity assessment

Alfonso Medela,[1] Alberto Sabater[1] Ignacio Hernández Montilla,[1] Taig Mac Carthy,[2] Andy Aguilar,[2] Gerardo Fernández,[1] Antonio Martorell,[2] Sofia Vera Carretero,[1] Constanza Balboni,[3] Javier Martín Alcalde,[3] Jose Luis Ramírez Bellver,[4] Alejandro Martin-Gorgojo,[5] Pedro Rodriguez Jiménez[6] and Jose Luis López Estebaranz[3]

[1]Department of Medical Data Science, Legit.Health, Bilbao, Spain
[2]Department of Clinical Endpoint Innovation, Legit.Health, Bilbao, Spain
[3]Department of Dermatology, DermoMedic Center, Hospital Universitario Fundación de Alcorcón, Madrid, Spain
[4]Dermatology Department, Instituto Médico Ricart, Madrid, Spain
[5]STI/Dermatology Department, Madrid City Council, Madrid, Spain
[6]Department of Dermatology, Hospital Universitario de la Princesa, Hospital Ruber Internacional, Madrid, Spain
Correspondence: Alfonso Medela. Email: alfonso@legit.health

A. Medela and A. Sabater contributed equally to this work.

Abstract

Background **Acne affects approximately 80% of adolescents worldwide, regardless of age, sex or ethnicity. Effective treatment depends on timely and accurate severity assessment. However, the lack of methodological standardization results in high inter- and intraobserver variability. A standardized and explainable grading tool is needed to support consistent acne severity evaluation.**

Objectives **To develop an artificial intelligence-driven tool that quantifies inflammatory lesion count and density from facial images to calculate scores aligned with the Investigator Global Assessment (IGA) scale.**

Methods **ALADIN (Acne Lesion And Density INdex) is designed as a simple mathematical model able to deliver reproducible and interpretable acne severity scores. We validated ALADIN retrospectively using two datasets: (i) a public dataset with predominantly severe cases of acne and (ii) a private dataset of mild-to-moderate cases. Validation results evaluated ALADIN against IGA and Global Acne Grading System (GAGS) scores, annotated by expert dermatologists.**

Results **Lesion detection with ALADIN was evaluated using precision, recall, mean average precision and mean absolute error (MAE). Acne grading performance was assessed via Pearson/Spearman correlations, MAE and Cohen's kappa. ALADIN's performance aligned well with the interobserver correlation of dermatologists, achieving Pearson correlations of 0.58 vs. 0.56 and 0.65 vs. 0.65 (private and public datasets, respectively) for IGA scores, and 0.32 vs. 0.44 and 0.43 vs. 0.44 for GAGS scores.**

Conclusions **ALADIN demonstrates consistent performance and excellent interpretability. Compared with manual scoring systems, it provides a streamlined, data-driven alternative capable of minimizing observer bias. When integrated into a medical device software, it can further aid in standardizing acne severity assessment and enhance patient monitoring. However, in this study, analyses were predominately carried out with White and Asian patients. Future work will validate and ensure robustness with darker skin tones.**

What is already known about this topic?

- Assessing acne severity is crucial for patient care and clinical trials, but standard methods are subjective or overly complex.
- This leads to significant variability between dermatologists, complicating treatment and research.
- While automated systems have been explored to provide objectivity, there is still a need for a reliable and interpretable tool that aligns with clinical judgement and can be easily integrated into medical practice to reduce this inconsistency.

Accepted: 25 March 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of British Association of Dermatologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

What does this study add?

- This study introduces ALADIN (Acne Lesion And Density INdex), an automated tool that quantifies acne severity from images by analysing lesion count and density.
- ALADIN provides a simple, explainable formula optimized to replicate the clinical judgement of the Investigator Global Assessment scale.
- This objective score is designed to reduce interobserver variability among dermatologists, standardizing patient assessment for clinical practice and research, and acting as a supportive tool to enhance, not replace, clinical judgement.

Acne is an inflammatory skin disease caused by dysfunction in the sebaceous glands, leading to facial lesions such as blackheads, whiteheads, pimples and scars.[1,2] Its development is largely driven by androgens, which increase oil production, especially during adolescence and early adulthood.[3] Beyond its physical symptoms, acne often causes emotional distress, including low self-confidence, social anxiety and depression.[1,4] It is one of the most common skin disorders globally, affecting around 80% of teenagers.[1,3] Acne's unpredictable flare-ups make treatment difficult.[5] While severity grading is essential, its polymorphic nature leads to inconsistent assessment criteria among dermatologists.[6] More than 25 grading systems exist,[2] creating barriers to consistent care and research standardization.[7,8]

Lesion counting is among the most objective methods for assessing acne severity. A key example is the Hayashi critera,[9] which categorize acne into five severity levels based on facial lesion count. However, lesion counting is time-consuming,[8] does not account for factors such as lesion concentration or size,[7] and is subject to high interobserver variability. Moreover, acne lesion categorization is challenging, especially when assessed from images. As alternatives, two methodologies have gained popularity. Firstly, the Investigator Global Assessment (IGA)[10] uses broad and subjective criteria, making it time-efficient but prone to inter- and intraobserver variability. Secondly, the Global Acne Grading System (GAGS),[11,12] which is based on more objective guidelines, but is overly complex, time-consuming and overlooks lesion concentration.

Despite recent advancements in artificial intelligence-driven acne severity assessment,[13,14] existing methods frequently suffer from limited explainability (black boxes) or poor generalization, or overlook acne attributes such as lesion size and concentration. While these approaches have demonstrated technical promise, a significant gap remains for a tool that balances the objectivity of quantitative analysis with the interpretability required for clinical utility.

To address these limitations, we introduce ALADIN (Acne Lesion And Density INdex), a novel, automated system designed to provide an objective, reproducible and transparent proxy for the IGA. ALADIN is not merely another lesion counter or a black-box classifier. Instead, it is built on the clinical hypothesis that a dermatologist's global assessment is primarily informed by two key visual cues: the total number of inflammatory lesions and their spatial density. By using computer vision to precisely quantify these two features from a standard digital image and integrating them into a simple, explainable formula, ALADIN aims to distil the essence of expert clinical reasoning into a consistent score. This study details the development of the ALADIN formula, validates its performance against expert dermatologist ratings on two distinct datasets, and discusses its potential to reduce interobserver variability and serve as a standardized endpoint for clinical management and research trials. Moreover, we integrate ALADIN into the Legit.Health medical device to enhance its potential for helping to standardize acne assessment protocols and improve patient management and clinical trial consistency.

Materials and methods

Acne severity scoring systems

Of the numerous acne severity scoring systems, the IGA and the GAGS stand out for their simplicity and objectivity, respectively.

Investigator Global Assessment

The IGA is a scoring system that rates a pathology on a scale from 0 to 4 using nonrigid guidelines that appeal to the occurrence of noninflammatory, inflammatory and nodular lesions. As the IGA lacks objective guidelines, it allows not only for broader analysis of the affected area, but is also prone to high variability between clinics, trials and doctors.

Global Acne Grading System

The GAGS quantifies acne severity by assessing four lesion types (comedones, papules, pustules, nodules/cysts) across six locations (forehead, cheeks, nose, chin, chest, back). Each location receives a severity score (1–4) based on its most severe lesion, which is then multiplied by a predefined weighting factor (1–3), reflecting its surface area and pilosebaceous unit density, yielding a total score from 0 to 52. For the ALADIN system, which focuses on facial images, GAGS was adapted by excluding the chest and back (reducing the maximum score to 32) and introducing a 'Not visible' option for facial regions not fully depicted in the image, treating these areas as having 'No lesions'.

Datasets and study design

This retrospective noninterventional study involved three different datasets for the optimization and evaluation of the two main components of this work: the AcneDetectorModel and the ALADIN endpoint. The three datasets include images of patients with facial acne.

AcneDetection

The AcneDetection dataset consists of 2755 images featuring individuals with facial acne and images of people with no visible acne lesions, used for the training of the AcneDetectorModel. Acne-related images are annotated with bounding boxes for inflammatory lesions, which only reflect the location of a lesion, not the kind of lesion (e.g. pustules and papules). The dataset includes 1204 images sourced from the ACNE04 dataset (with a minimum of 1, a

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

maximum of 434 and a mean of 27 annotations per image)[15,16] and 894 images sourced from RoboFlow (https://roboflow.com; minimum of 0, maximum of 122 a mean of 10 annotations per image). To reduce false-positive results, we incorporated 75 images from the SFA dataset and 564 images from the Describable Textures Dataset (DTD-Textures),[17,18] with no bounding box annotations. The dataset predominantly features White and Asian individuals (male, $n=700$; female, $n=1398$; close-up image with unrecognizable gender, $n=93$; Fitzpatrick skin phototypes I and II, $n=1055$; phototypes III and IV, $n=1114$; phototypes V and VI, $n=22$). The AcneDetection dataset is split into train and validation splits with 2125 and 630 samples, respectively.

AcneSeverity-V1 dataset

The AcneSeverity-V1 dataset consists of 331 high-resolution images provided by Clínica Dermatológica Internacional, Spain. Images are taken in clinical practice under controlled illumination conditions with White individuals (male, $n=66$; female, $n=265$; Fitzpatrick skin phototypes I and II, $n=282$; phototypes III and IV, $n=49$) with predominantly mild-to-moderate facial acne. All patients provided written informed consent to be included. To reduce sensitive data collection to a minimum, no more demographic information has been collected. The images are labelled using IGA (IGA 0 = 14, IGA 1 = 65, IGA 2 = 150, IGA 3 = 83, IGA 4 = 19) and GAGS scores (minimum score 0, maximum score 20, mean score 6) by three expert dermatologists (P.R.J., A.M.G, J.L.R.B.). This dataset is intended for the design of the ALADIN mathematical expression and its evaluation. As ALADIN is a simple mathematical expression optimized via symbolic learning rather than a deep neural network, it does not require a conventional train/validation split. This approach is considerably less data-dependent than deep learning models, making the 331-image dataset sufficient for robustly tuning the limited number of parameters in the ALADIN formula.

AcneSeverity-V2 dataset

The AcneSeverity-V2 dataset is composed of 27 images collected from public dermatology atlases, including DermAtlas (https://www.dermatlas.net/) and Atlas Dermatológico (https://www.atlasdermatologico.com.br/). Selected images have a minimum resolution of 640 pixels on their longer side and display at least two-thirds of a face side, excluding highly localized facial areas. The dataset primarily features White individuals with severe acne (male, $n=9$; female, $n=18$; Fitzpatrick skin phototypes I and II, $n=13$; phototypes III and IV, $n=12$; and phototypes V and VI, $n=2$). These images were annotated by two expert dermatologists (C.B., J.M.A.) with IGA (IGA 0 = 1, IGA 1 = 3, IGA 2 = 10, IGA 3 = 11, IGA 4 = 2) and GAGS scores (minimum 2, maximum 28, mean 11). Given the number of annotators and its limited size, the AcneSeverity-V2 dataset is intended as an exploratory test set for evaluating ALADIN's generalization, not for its design. While the small sample size warrants caution in interpreting the results, this dataset provides valuable insights into the model's performance on these third-party cases of severe acne, which were less prevalent in the V1 dataset.

For the development and validation of ALADIN, the IGA scores from the AcneSeverity-V1 and AcneSeverity-V2 datasets were used as the ground truth. These reference values were established through the mathematical consensus between the expert annotators, ensuring that each score represents a robust clinical assessment of acne severity for the corresponding image. Similarly, a consensus for the GAGS scores was also calculated for further validation.

Evaluation metrics and benchmarks

We evaluate the performance of the lesion detection models with common object detection metrics: precision (P), recall (R) and mean average precision (mAP@50). Then, we evaluate the lesion-counting performance with the mean absolute error (MAE).

Acne severity assessment performance is evaluated using the IGA and GAGS scores from the AcneSeverity-V1 and AcneSeverity-V2 datasets. Firstly, we evaluate the alignment between ALADIN and the annotators with respect to the ground-truth IGA scores. Secondly, we evaluate the alignment between ALADIN and the annotators with respect to the ground-truth GAGS scores. Evaluation metrics include the Spearman and Pearson correlations, MAE and Cohen's kappa coefficient with quadratic weights (CK). Correlation metrics are applied both to IGA and GAGS scores, while MAE and CK are computed only for IGA evaluation, whose scale aligns with ALADIN's. Note that, as the AcneSeverity-V1 and AcneSeverity-V2 datasets are not annotated with lesion counts, we do not use them to evaluate the performance of lesion counting.

ALADIN

Acne severity grading is frequently reduced to lesion counting.[9] However, lesion concentration is also an important clinical indicator for diagnosis and severity grading.[19] The proposed ALADIN scoring system combines lesion number (N) and lesion density (D) into a single and easy-to-interpret mathematical formula, providing a clear and accessible approach to acne severity grading:

$$ALADIN = N^a \cdot (D + b) \qquad \text{(Equation 1)}$$

This formula was developed using a two-stage semiautomatic process based on symbolic learning and differential evolution optimization.[20,21] The symbolic learning stage generates a set of mathematical expressions that combine N and D with different mathematical operands. In the second stage, we add constant variables (a and b weights), when possible, to the mathematical expressions and refine them with differential evolution optimization. This entire optimization process is designed to maximize Cohen's kappa score by aligning ALADIN's output with the IGA scores from the AcneSeverity-V1 dataset. As a result of this process, we set ALADIN as the formula (see Equation 1) with the best trade-off between performance and interpretability. More complex formulas offered marginal performance gains but were less clinically interpretable. As shown in Figure 1, ALADIN scores grow with the number of lesions and density and are finally clipped to the 0–4 range to align with the IGA scale.

For the symbolic learning stage, we used the PySR Python library with 250 iterations,[20] 50 populations, population size 50, maximum size 10 and L2 loss. For the differential evolution optimization stage, we use the SciPy Python library for 10 000 iterations (remaining hyperparameters set as default).[22]

Lesion counting

To automate lesion counting, we have trained a deep learning model: the AcneDetectorModel. Given a simple smartphone

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

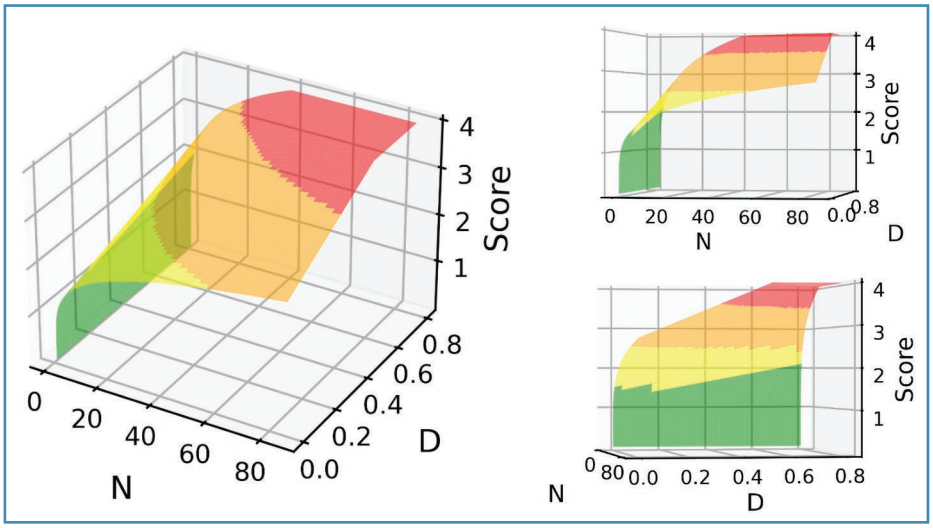

Figure 1 ALADIN (Acne Lesion And Density INdex) scoring space from different angles. The bottom axes refer to lesion number (*N*) and lesion density (*D*) values, and the vertical axis refers to the calculated ALADIN scores. Results are cut off at a score of 4 to align with the Investigator Global Assessment (IGA) scoring system. A score of 0 corresponds to individuals with no lesions and a score of 4 to those with a high number and density of lesions. Colours show the different IGA discrete scores.

image, the AcneDetectorModel effectively localizes and dimensionalizes inflammatory acne lesions, enabling straightforward lesion counting. The model was intentionally trained to detect inflammatory lesions as a single, unified class. This design choice is rooted in the fact that the clinical differentiation between lesion subtypes (e.g. papules, pustules and nodules) is a well-documented source of high interobserver variability in acne grading,[23] minimizing the subjectivity in the data annotation and allowing for a more reliable acne lesion detector.

The AcneDetectorModel is based on the YOLOv11-M architecture (https://github.com/ultralytics/ultralytics), which we fine-tuned on the AcneDetection dataset with cross-validation. This single-stage detection architecture balances high performance with computational efficiency. Of special relevance, unlike traditional computer vision models, which commonly resize images to 224–384 pixels, we resize input images so that their longer dimension is 640 pixels. This approach preserves the visibility of small acne lesions, ensuring they retain sufficient detail for accurate detection. We train the model for 95 epochs, with batch size 64, AdamW optimizer, learning rate of 5×10^{-4}, cosine learning rate annealing, and standard geometric and light data augmentations.

Density calculation

Given the localization and dimensions of detected acne lesions, we define acne lesion density as the ratio between the overlapping detection area and the total area covered by detected lesions (Figure 2c). This density score is 0- to 1-bounded and accounts for lesion location, concentration and size.

Prior to computing density, we preprocess detections by first converting the detected bounding boxes into circles and enlarging their radius (Figure 2a, b). The circular transformation refines lesion boundaries by excluding irrelevant bounding box corners, while enlargement improves collision detection among closely positioned lesions. To determine the optimal enlargement factor, we performed a parameter sweep, testing integer values from 2 to 10. The factor of 6 was selected as it yielded the highest alignment between the final ALADIN score and the ground-truth IGA scores.

Legit.Health

ALADIN was designed for its integration into a Class IIb, CE-marked medical device called Legit.Health (ISO approval number: 13485; certificate number: MDR 792790; SRN: ES-MF-000025345). This device is accessible from computers and smartphones through a web application, to ease its use by clinicians and patients. Similar to ALADIN, Legit.Health offers a broad set of tools for automating clinical endpoints. These include other scoring systems such as SCORing Atopic Dermatitis, Psoriasis Area and Severity Index or the Automatic International Hidradenitis Suppurativa Severity Score System; a skin diagnosis support tool; and image domain and quality-check tools to guarantee precise image analysis results.

To assess acne severity, the Legit.Health medical device follows three stages. Firstly, users take and upload an image of the face, which can be manually cropped to focus on the area of interest. These images are then stored in the computer-aided diagnosis system. Secondly, the ALADIN framework analyses the uploaded image and automatically determines the severity of acne. Finally, the system generates a user-friendly report that includes the analysed image with the detected lesions, the total number of lesions and density, and a chart illustrating the severity level and its progression over time (Figure 3). Unlike the IGA, which provides discrete values, ALADIN produces continuous severity scores. To facilitate the severity interpretation and its evolution over time, in the Legit.Health integration, ALADIN's output is scaled by a factor $\frac{5}{2}$, ensuring alignment with a 0–10 scale.

Results

We evaluated the proposed AcneDetectorModel for the task of acne lesion detection and counting, and the ALADIN endpoint for the task of acne severity assessment.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

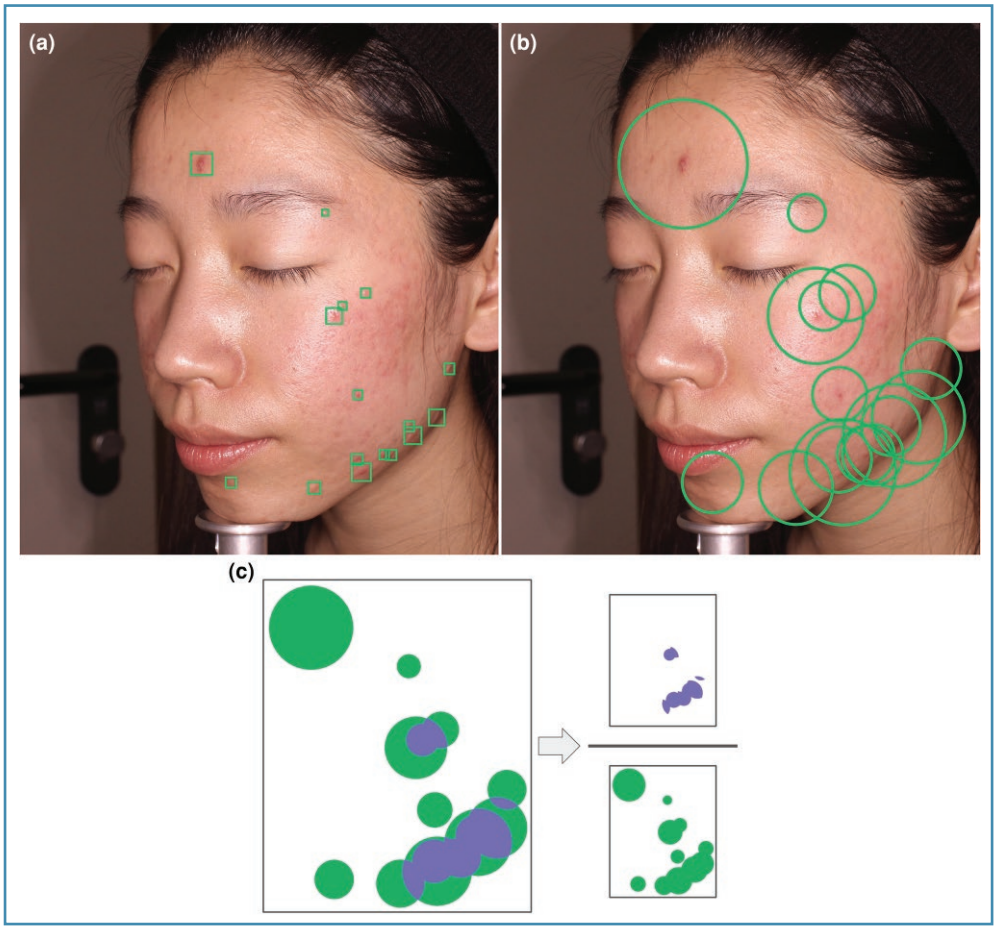

Figure 2 Acne lesion density calculator. Original detections (a) are transformed into circles and then (b) enlarged. (c) Final density (D) is calculated as the ratio between the area with collisions and the area covered by all the detections.

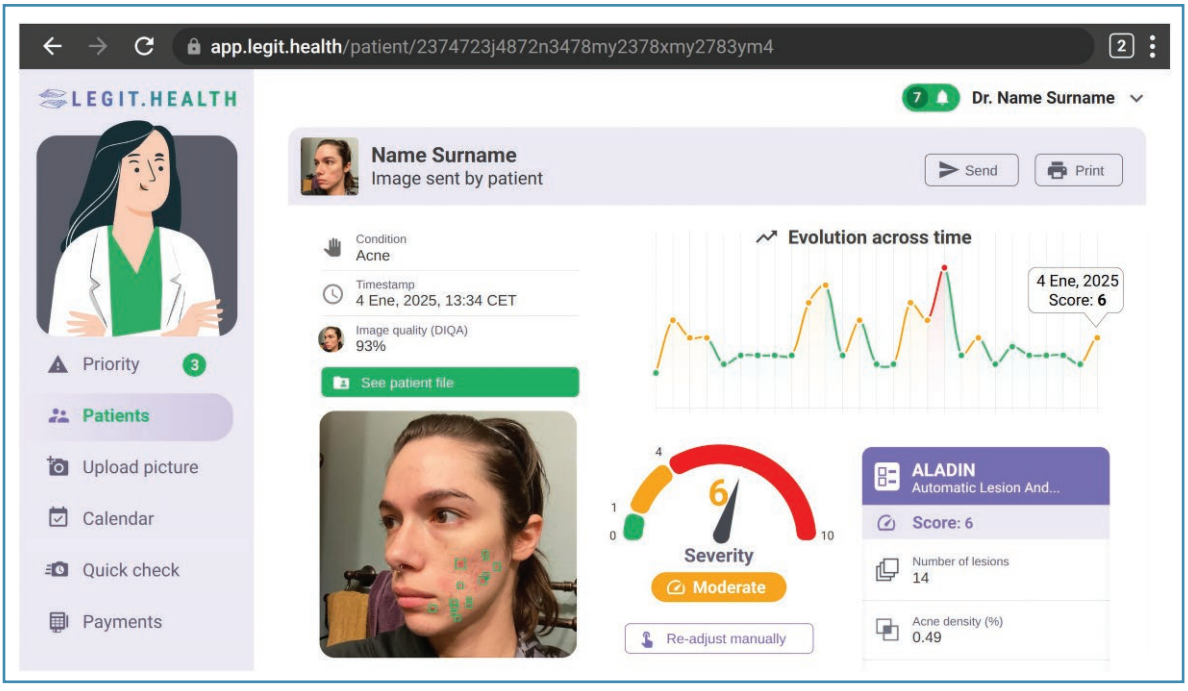

Figure 3 The integration of ALADIN (Acne Lesion And Density INdex) into the Legit.Health medical device. The generated report includes information on the current and past acne severity.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

Automatic lesion counting

For the acne lesion detection task, we trained and evaluated a family of YOLOv11 models [Nano (N), Small (S) and Medium (M)] on the AcneDetection dataset using cross-validation. The trained models showed comparable MAE (7.02, 7.05 and 7.15, respectively) for the task of lesion counting. However, they differed in terms of precision (0.44, 0.49 and 0.50, respectively), recall (0.36, 0.40 and 0.41, respectively) and mAP (0.34, 0.40 and 0.42, respectively). Among the models, YOLOv11 (M) outperformed the others. Consequently, for the following evaluation and ALADIN endpoint development, we selected YOLOv11 (M) as the AcneDetectorModel.

ALADIN performance

To evaluate the performance of the proposed acne severity scoring system, we compute ALADIN scores for the images from the AcneSeverity-V1 and AcneSeverity-V2 datasets. These scores are calculated using the number of lesions and density values calculated from AcneDetectorModel detections.

As shown in Table 1, ALADIN achieves IGA performance statistics aligned to the interobserver agreement among expert dermatologists on the AcneSeverity-V1 dataset. Notably, ALADIN obtained a higher Cohen's kappa score [0.53 vs. 0.46 (ALADIN vs. dermatologists)], while correlations (Pearson 0.58 vs. 0.56 and Spearman 0.56 vs. 0.58) remained similar. On the AcneSeverity-V2 dataset, ALADIN demonstrates strong generalization, with a correlation (Pearson 0.65 vs. 0.65 and Spearman 0.54 vs. 0.58) and Cohen's Kappa (0.61 vs. 0.62) aligned with that of expert dermatologists. Moreover, the lower MAE statistics [0.63 (0.49) and 0.64 (0.71) in both datasets, respectively] highlight that even when ALADIN fails, it does so in the vicinity of the ground-truth score. These IGA performance results show ALADIN's capability of automating the acne severity assessment, reducing interobserver variability and contributing to a more standardized acne severity scoring system.

Regarding the correlation with GAGS scores (Table 2), correlation between IGA and GAGS scores was expectedly lower and comparable between ALADIN (Pearson 0.32 and 0.43, and Spearman 0.34 and 0.52 in the AcneSeverity-V1 and AcneSeverity-V2 datasets, respectively) and the dermatologists (Pearson 0.44 and 0.44, and Spearman 0.44 and 0.33, respectively). This is presumably due to two factors. Firstly, the ALADIN design uses a more restricted scale (0–4, like the IGA) than GAGS (0–32). Secondly, ALADIN and IGA emphasize lesion number and density, whereas GAGS focuses solely on lesion type, disregarding quantitative aspects. This highlights the weak relationship between IGA and GAGS, stemming from their divergent evaluation frameworks rather than deficiencies in ALADIN.

Discussion

Artificial intelligence has been shown to be highly effective in grading dermatological conditions.[24,25] For acne, current computer vision methods predominantly include classification models and lesion-counting algorithms.[13,14,26] While classifiers provide rapid severity assessments, they often are black boxes that lack explainability and generalizability. Counting approaches improve objectivity but miss critical details like lesion size and distribution. ALADIN advances this field by integrating lesion count, size and density, delivering a more interpretable and precise scoring system. When integrated into the Legit.Health software, ALADIN works in synergy with tools such as automated image quality and diagnostic support tools, enabling clinicians to conduct a more informed severity assessment, reducing inter- and intraobserver variability, and enhancing patient management and clinical trial standardization.

Analysis of IGA evaluation annotators' performance revealed certain interobserver variability, even with dermatologists from the same clinic, largely due to the scale's vague descriptors (e.g. 'some' and 'many') and reliance on clinical individual experiences. In the AcneSeverity-V1 dataset, this led to frequent misclassification, such as mistaking postinflammatory erythema for active acne or missing low-severity lesions. However, ALADIN occasionally misclassifies 0 and 1 scores by detecting lesions in low acne-affected images that the annotators did not. Despite these challenges, ALADIN's performance was aligned with that of clinicians. Notably, in the AcneSeverity-V1 dataset, it achieved a higher Cohen's kappa score (0.53 vs.0.46), underscoring its potential for clinical reliability.

Although GAGS offers a more structured grading system, it focuses solely on the most severe lesions, disregarding milder ones, their frequency and their localized distribution.[11,12] As a result of the incompatibility between IGA and GAGS scores, although ALADIN still performs on a par with expert dermatologists, both exhibit a lower correlation. To bridge the gap with the GAGS framework, it would be necessary to (i) compile a more comprehensive dataset, (ii) design a multiclass lesion detector and (iii) design a more complex severity mathematical model. The new dataset must include images from all GAGS body regions and be labelled with lesion categories, which will likely increase the annotation variability. The new detector would shift to a multiclass system capable of distinguishing lesion subtypes (e.g. papules vs. nodules). The new mathematical model would include information about the different lesion categories and fine-grained localization. As a consequence, this new framework would hinder the exceptional interpretability of the proposed ALADIN solution and would be subject to increased annotation variability. Therefore, such an endpoint is outside the scope of this study.

Table 1 Performance of ALADIN (Acne Lesion And Density INdex) with regard to Investigator Global Assessment scores

	AcneSeverity-V1				AcneSeverity-V2			
	Spearman	Pearson	MAE	CK	Spearman	Pearson	MAE, mean (SD)	CK
ALADIN	0.56	0.58	0.63 (0.49)	0.53	0.54	0.65	0.64 (0.71)	0.61
Dermatologists	0.58	0.56	0.65 (0.63)	0.46	0.58	0.65	0.44 (0.50)	0.62

CK, Cohen's kappa; MAE, mean absolute error.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

Table 2 Performance of ALADIN (Acne Lesion And Density INdex) with regard to Global Acne Grading System scores

	AcneSeverity-V1		AcneSeverity-V2	
	Spearman	Pearson	Spearman	Pearson
ALADIN	0.34	0.32	0.52	0.43
Dermatologists	0.44	0.44	0.33	0.44

The core methodological strength of ALADIN is its focus on the count and density of inflammatory lesions, rather than attempting to subcategorize them. While systems like GAGS, which classify lesion subtypes, appear to be more granular, this high complexity is what introduces significant interobserver variability and subjectivity, undermining the goal of a standardized assessment.[23] Our results, which show a strong correlation between ALADIN and expert-assigned IGA scores, validate the hypothesis that an accurate and reliable global severity assessment does not depend on lesion subcategorization.

Despite ALADIN's strong performance, several limitations should be considered. As with any computer vision system, it is sensitive to image quality. High-resolution, well-lit and motion-free images are essential, especially given the small size of acne lesions. Accurate assessment also depends on capturing the most affected facial areas with images that, following the Hayashi criteria,[9] should cover the full face at approximately 70-degree angles to ensure consistency.

ALADIN's performance also depends on the quality of training data, which is subject to the interobservability between annotators. Owing to the difficulty of this task, similar to other studies,[27] more annotators would enhance the data quality and therefore the performance in categorizing acne severity. Moreover, the training data is biased towards certain acne severities. Counting with more low- and high-severity cases would enhance ALADIN's performance in these diverse scenarios. Lastly, this study primarily focused on White and Asian individuals, with lower representation of people with more richly pigmented skin types. Consequently, future work is essential to validate ALADIN's performance and ensure its robustness for patients with darker skin, addressing the potential variation in lesion detection challenges across diverse skin types. Such research will also involve multicentre data and a wider spectrum of acne severities and patient demographics.

Author contributions

Alfonso Medela (Conceptualization, Project administration [equal], Investigation [supporting]), Alberto Sabater (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing [lead]), Ignacio Hernandez Montilla (Data curation, Formal analysis, Investigation, Methodology, Visualization [equal], Writing—original draft [supporting]), Taig Mac Carthy (Conceptualization, Visualization [equal]), Andy Aguilar (Funding acquisition, Project administration [equal]), Gerardo Fernández (Project administration, Resources [equal]), Antonio Martorell (Writing—review & editing [equal]), Sofia Vera Carretero (Data curation, Investigation, Writing—review & editing [equal]), Constanza Balboni (Writing—review & editing [equal]), Javier Martín Alcalde (Writing—review & editing [equal]), Jose Luis Ramírez Bellver (Writing—review & editing [equal]), Alejandro Martin-Gorgojo (Writing—review & editing [equal]), Pedro Rodriguez Jiménez (Writing—review & editing [equal]), and Jose Luis Lopez Estebaranz (Writing—review & editing [equal]).

Funding sources

The authors acknowledge financial support provided by the European Union NextGenerationEU and the Centro para el Desarrollo Tecnológico y la Innovación (CDTI) through the support program PYME con Sello de Excelencia del Acelerador del EIC de Horizonte Europa (SoE-20221020).

Conflicts of interest

A.Mendela, T.M., A.A., G.F. and A.Martorell hold shares in AI LABS GROUP SL, while A.S. and I.H.M. are employed by the company. A.Martorell has received honoraria and/or travel grants and/or has served as a member of advisory committees for AbbVie, Aceleryn, Almirall, Amgen, Boehringer Ingelheim, Bristol-Myers Squibb, Celgene, Incyte, Johnson & Johnson, LEO Pharma, MoonLake, Novartis, Pfizer, Sandoz, Sanofi, UCB and Legit Health. He has also served as a principal investigator in clinical trials sponsored by AbbVie, Aceleryn, Alumis Almirall, Amgen, Boehringer Ingelheim, Bristol-Myers Squibb, Celgene, Incyte, Johnson & Johnson, LEO Pharma, MoonLake, Novartis, Pfizer, MSD, Takeda, Sanofi, UCB and MSD. J.L.E. served as a consultant, participated in clinical trials and/or received speaking fees from Almirall, BMS, Janssen, LEO Pharma, Lilly, AbbVie, Bioderma, Galderma, Glenmark, UCB, Legit. Health, Novartis, Pierre-Fabre, Invasix, Isdin and Incyte.

Data availability

Images used in the AcneDetection and AcneSeverity-V2 datasets were sourced from publicly available repositories, including ACNE04,[15,16] RoboFlow (https://roboflow.com), SFA,[17] Describable Textures Dataset (DTD-Textures),[18] DermAtlas (https://www.dermatlas.net/) and Atlas Dermatológico (https://www.atlasdermatologico.com.br/). The AcneSeverity-V1 is not publicly available due to specific institutional requirements governing privacy protection.

Ethics statement

This study involves the analysis of images compiled from online publicly available sources and images compiled from private sources. Publicly available images were used in accordance with the respective source policies, and the authors complied with General Data Protection Regulation and relevant local regulations. Private images were obtained from Clínica Dermatológica Internacional, with informed consent provided by all participants. The image collection was noninvasive, involved no treatment or drug administration, and posed no risk to the participants. The research was conducted in accordance with the STARD guidelines.

Patient consent

Figures that include human participants were extracted from the public datasets ACNE04 and ACNE_KAGGLE. Private images were obtained from Clínica Dermatológica Internacional, with informed consent provided by all participants.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026

References

1 Ak M. A comprehensive review of acne vulgaris. *Clin Pharm* 2019; **1**:17–45.
2 Heng AH, Chew FT. Systematic review of the epidemiology of acne vulgaris. *Sci Rep* 2020; **10**:5754.
3 Pérez Armas R, Calaña Ramírez A. Acne polimorfo: tratamiento con implacen. *Rev Cuba Med Gen Integral* 1995; **11**:285–90.
4 Zaenglein AL. Acne vulgaris. *N Engl J Med* 2018; **379**:1343–52.
5 Layton AM. The Leeds Acne Grading Technique. In: *Pathogenesis and Treatment of Acne and Rosacea* (Zouboulis CC, Katsambas AD, Kligman AM, eds). Berlin, Heidelberg: Springer, 2014; 317–24.
6 Adityan B, Kumari R, Thappa DM. Scoring systems in acne vulgaris. *Indian J Dermatol Venereol Leprol* 2009; **75**:323.
7 Agnew T, Furber G, Leach M, Segal L. A comprehensive critique and review of published measures of acne severity. *J Clin Aesthet Dermatol* 2016; **9**:40.
8 Alsulaimani H, Kokandi A, Khawandanh S, Hamad R. Severity of acne vulgaris: comparison of two assessment methods. *Clin Cosmet Investig Dermatol* 2020; **13**:711–16.
9 Hayashi N, Akamatsu H, Kawashima M, Acne Study Group. Establishment of grading criteria for acne severity. *J Dermatol* 2008; **35**:255–60.
10 US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Acne Vulgaris: establishing effectiveness of drugs intended for treatment. Available at: https://www.fda.gov/media/71152/download (last accessed 17 March 2026).
11 Doshi A, Zaheer A, Stiller MJ. A comparison of current acne grading systems and proposal of a novel system. *Int J Dermatol* 1997; **36**:416–18.
12 Zohra FT, Sultana T, Islam S, Nasreen T. Evaluation of severity in patients of acne vulgaris by global acne grading system in Bangladesh. *Clin Pathol* 2017; **1**:000105.
13 Junayed MS, Jeny AA, Atik ST *et al.* AcneNet – a deep CNN based classification approach for acne classes. Available at: https://ieeexplore.ieee.org/document/8850935 (last accessed 17 March 2026).
14 Rashataprucksa K, Chuangchaichatchavarn C, Triukose S *et al.* Acne detection with deep neural networks. Available at: https://dl.acm.org/doi/epdf/10.1145/3421558.3421566 (last accessed 17 March 2026).
15 Wu X, Wen N, Liang J *et al.* Joint acne image grading and counting via label distribution learning. Available at: https://ieeexplore.ieee.org/document/9010021 (last accessed 17 March 2026).
16 Gazeau L, Nguyen H, Nguyen Z *et al.* AcneAI: a new acne severity assessment method using digital images and deep learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (pp. 68–78). Cham: Springer Nature Switzerland. Available at: https://papers.miccai.org/miccai-2024/042-Paper2216.html
17 Casati JP, Moraes DR, Rodrigues EL. *SFA: A Human Skin Image Database Based on FERET and AR Facial Images*. Rio de Janeiro: São Carlos School of Engineering, University of São Paulo, 2013.
18 Cimpoi M, Maji S, Kokkinos I *et al.* Describing textures in the wild. Available at: https://ieeexplore.ieee.org/document/6909856 (last accessed 17 March 2026).
19 Tan J, Kircik L. Evolution of acne assessments and impact on acne medications: an evolving, imperfect paradigm. *J Drugs Dermatol* 2016; **15**:79–86.
20 Cranmer M. Interpretable machine learning for science with PySR and Symbolic Regression.jl. Available at: https://arxiv.org/abs/2305.01582 (last accessed 17 March 2026).
21 Storn R, Price K. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 1997; **11**:341–59.
22 Virtanen P, Gommers R, Oliphant TE *et al.* Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**:261–72.
23 Tanghetti EA, Kircik LH, Green LJ *et al.* A phase 2, multicenter, double-blind, randomized, vehicle-controlled clinical study to compare the safety and efficacy of a novel tazarotene 0.045% lotion and tazarotene 0.1% cream in the treatment of moderate-to-severe acne vulgaris. *J Drugs Dermatol* 2019; **18**:542.
24 Huang K, Wu X, Li Y *et al.* Artificial intelligence-based psoriasis severity assessment: real-world study and application. *J Med Internet Res* 2023; **25**:e44932.
25 Hernández Montilla I, Medela A, Mac Carthy T *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.
26 Maroni G, Ermidoro M, Previdi F, Bigini G. Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity. Available at: https://ieeexplore.ieee.org/document/8280925 (last accessed 17 March 2026).
27 Alsulaimani H, Kokandi A, Khawandanh S, Hamad R. Severity of acne vulgaris: comparison of two assessment methods. *Clin Cosmet Investig Dermatol* 2020; **28**:711–16.

Downloaded from https://academic.oup.com/skinhd/advance-article/doi/10.1093/skinhd/vzag062/8704397 by guest on 15 June 2026