Received: 13 February 2023 | Accepted: 13 May 2023

DOI: 10.1111/srt.13357

ORIGINAL ARTICLE

WILEY

Automatic International Hidradenitis Suppurativa Severity Score System (AIHS4): A novel tool to assess the severity of hidradenitis suppurativa using artificial intelligence

Ignacio Hernández Montilla[1] | Alfonso Medela[1] | Taig Mac Carthy[2,#] | Andy Aguilar[2] | Pedro Gómez Tejerina[1] | Alejandro Vilas Sueiro[3] | Ana María González Pérez[4] | Laura Vergara de la Campa[5] | Loreto Luna Bastante[6] | Rubén García Castro[7] | Fernando Alfageme Roldán[8,#]

[1]Department of Medical Computer Vision and PROMs, LEGIT.HEALTH, Bilbao, Spain

[2]Department of Clinical Endpoint Innovation, LEGIT.HEALTH, Bilbao, Spain

[3]Dermatology Unit, Ferrol Teaching University Hospital Complex, Ferrol, Spain

[4]Dermatology Unit, Salamanca Teaching University Hospital, Zamora, Spain

[5]Dermatology Unit, Toledo Teaching University Hospital Complex, Toledo, Spain

[6]Dermatology Unit, Rey Juan Carlos Teaching University Hospital, Madrid, Spain

[7]Dermatology Unit, Fundacion Jiménez Díaz Teaching University Hospital, Madrid, Spain

[8]Dermatology Unit, Puerta de Hierro Hospital, Madrid, Spain

Correspondence
Ignacio Hernández Montilla, Department of Medical Computer Vision and PROMs, Legit.Health, 48013, Bilbao, Spain.
Email: ignaciohernandez@legit.health

Funding information
Department of Economic Development and Infrastructures of the Basque Government (HAZITEK Program); European Regional Development Fund (ERDF)

Abstract

Background: Hidradenitis suppurativa (HS) is a painful chronic inflammatory skin disease that affects up to 4% of the European adult population. International Hidradenitis Suppurativa Severity Score System (IHS4) is a dynamic scoring tool that was developed to be incorporated into the doctor's daily clinical practice and clinical studies. This helps measure disease severity and guides the therapeutic strategy. However, IHS4 assessment is a time-consuming and manual process, with high inter-observer variability and high dependence on the observer's expertise.

Materials and methods: We introduce the Automatic International Hidradenitis Suppurativa Severity Score System (AIHS4), an automatic equivalent of IHS4 that deploys a deep learning model for lesion detection, called Legit.Health-IHS4net, based on the YOLOv5 architecture. AIHS4 was trained on Legit.Health-HS-IHS4, a collection of HS images manually annotated by six specialists and processed by a novel knowledge unification algorithm.

Results: Our results show that, with the current dataset size, our tool assesses the severity of HS cases with a performance comparable to that of the most expert physician. Furthermore, the model can be implemented into CADx systems to support doctors in their clinical practice and act as a new endpoint in clinical trials.

Conclusion: Our work proves the potential usefulness of artificial intelligence in the practice of evidence-based dermatology: models trained on the consensus of large clinical boards have the potential to empower dermatologists in their daily practice and replace current standard clinical endpoints.

[#]Taig Mac Carthy and Fernando Alfageme Roldán contributed equally to this work.

KEYWORDS
artificial intelligence, automatic severity assessment, CADx system, clinical decision support, hidradenitis suppurativa, IHS4

1 | INTRODUCTION

Hidradenitis suppurativa (HS) is a chronic, inflammatory, recurrent, painful, debilitating skin disease.[1,2] It manifests itself after puberty with inflamed lesions in apocrine gland-bearing areas of the body and leads to significantly impaired quality of life, depression, and handicap. It also exhibits many comorbidities like spondyloarthropathy, inflammatory bowel disease, obesity, and metabolic syndrome, which increase the burden of the disease.[3] HS is characterized by the formation of painful lesions, such as inflammatory nodules, abscesses, and pus-discharging tunnels, also known as fistulas and sinus tracts. These systematic manifestations appear typically, but not exclusively, in skin folds near armpits, groin, gluteal and perianal areas of the body—regions where apocrine glands are abundant.

HS affects around 1% of the general population, with some studies suggesting a prevalence up to 4%. It appears to be more common among females, although some authors have argued that certain locations have sexual predilection and might affect males more than females.[4] Although the etiology of the pathology is not fully understood, HS-related lesions appear due to the occlusion of hair follicles associated with apocrine glands.

Diagnosis of HS is usually performed through clinical observation. In other words: the assessment of the disease's severity is performed through manual scoring systems that are filled in subjectively. Due to the use of this outcome measure method, the expertise of the dermatologist plays a very significant role and its inter-observer variability is very high.

Several attempts have been made to standardize the assessment of the severity of HS by using different scoring systems, such as Modified Sartorius Score, Hurley classification, and Physician's Global Assessment. However, they are often too difficult to use in daily clinical practice and are generally poorly validated.[5]

Among the many scoring systems that can be used in clinical trials and in daily clinical practice, the International Hidradenitis Suppurativa Severity Score System (IHS4) is the most widely used by physicians.[6] The IHS4 is a validated tool that assigns a weighted score to lesions by dividing them into three categories: inflammatory nodules, abscesses, and draining tunnels. The score of the IHS4 is interpreted into qualitative meaning as "mild," "moderate," or "severe." This tool helps physicians assess the severity of the disease dynamically and can be used both in clinical research as well as daily clinical practice.

On the other hand, the field of dermatology is benefiting from recent advances in telemedicine. The reason is that dermatology is particularly suited for this healthcare model, due to its strong dependence on visual cues.[7] Indeed, the majority of skin disorders are visible to the naked eye, thus enabling smartphone cameras to collect clinical information remotely.

The visual component of dermatology has also made it susceptible to benefiting from breakthroughs of artificial intelligence in image processing. Convolutional Neural Networks (CNNs) can be trained with large clinical, dermatoscopic, and dermatopathological image databases to make predictions, which turns them into a promising clinical decision support tool that helps dermatologists in the diagnosis of a variety of disorders such as atopic dermatitis, psoriasis, onychomycosis, and melanoma.[8]

To support the results presented in this work, we first demonstrate the high inter-observer variability in the current HS severity assessment process. As a solution to the problem of inter-observer variability, we introduce the Automatic International Hidradenitis Suppurativa Severity Score System (AIHS4): the first AI-powered tool trained on the clinical consensus that automatically fills in the IHS4 scoring system. The AIHS4 reads images that can be taken with a regular phone camera and obtains the items of the IHS4 automatically. This could potentially reduce the time that physicians spend filling the pen-and-paper manual scoring systems, and improve the reliability of the outcome by reducing inter-observer variability. Finally, we show how the AIHS4 can be deployed into a CADx system that enables the use of the automatic version of this clinically-validated and widely used scoring system, thus empowering dermatologists in their daily practice, allowing the implementation of evidence-based dermatology and replacing current standard endpoints in clinical trials.

2 | MATERIALS AND METHODS

2.1 | Dataset and annotations

For this study, we created a dataset, called *Legit.Health-HS-IHS4*, which comprises 221 instances of the disease at different grades of severity. The dataset offers a wide range of perspectives and image sizes, as well as a significant diversity of skin tones. It also depicts lesions in different environments and situations: ranging from patches of skin up to entire body regions, with and without clothing. The dataset was created using a subset of the DermQuest and DermnetNZ datasets. The annotation was carried out by six specialists that frequently care for patients with HS. This resulted in six label sets per image. Regarding the clinical expertise of the annotators, one of the six specialists (3) is a senior dermatologist with decades of experience and the highest degree of specialization in HS.

Table 1 provides some extra information about the clinical experience of the specialists involved in the annotation process. We divide specialists into two sub-groups, based on their clinical expertise. Group 1 is comprised of all the specialists, whereas Group 2 excludes the senior dermatologist. The purpose of this division is to measure the

TABLE 1 Clinical survey for the IHS4 annotation task. The table presents the aggregated responses of all six dermatologists.

Question	Answer		
	Min	Max	Average ± Standard deviation
Clinical experience in treating hidradenitis suppurativa (HS) (in years)	3	7	4.50 ± 1.38
How often do you use IHS4 in your daily clinical practice? (1 = never, 10 = always)	7	10	8.17 ± 1.47
How difficult did you find the annotation task? (1 = very easy, 10 = very difficult)	4	9	6.50 ± 1.56

effect of a higher clinical experience over the annotation process. In this regard, all specialists were asked about the perceived difficulty of the annotation task.

Specialists annotated three types of lesions: abscesses, draining tunnels, and nodules. To perform such tasks, the specialists draw bounding boxes around each of the lesions as they perceived them to be. The specialists had no time limit to perform the task and could re-visit the annotation any number of times. Once the images were labeled, the corresponding IHS4 scores were calculated automatically from each of the specialists' annotations (Equation 2).

If a specialist did not find a type of lesion in an image, we accounted the corresponding item as "null," meaning that there is either no lesion at all or a minor lesion that is not of concern in the scoring of IHS4. In this regard, only two images of the entire dataset were considered to have no lesions at all by all six experts. These two images contained only mild lesions (papules) and were excluded from the study. Nevertheless, we did add them to the lesion detection task but with no bounding boxes associated. And most importantly, regarding the lesion detection task, most of the remaining 219 images were labeled differently by all six dermatologists.

The performance of each specialist at the image level is summarized in Table 2. This annotation summary, which is further explained later, reflects the inter-observer variability that occurs in real-life clinical practice. This inter-observer variability is particularly high in abscesses and nodules: the data show that specialists both under-detect and over-detect abscesses and nodules. We believe this is due to the differences in clinical expertise of the specialists, as well as the inability to physically examine that lesion during clinical assessment (i.e., not being able to palpate the lesion). This annotation variability poses a challenge when defining the best ground truth for the deep learning models in our object detection task.

2.2 | Ground truth labels

To the best of our knowledge, literature on methods for combining clinical knowledge in order to clear noisy annotations is sparse.[9,10] This poses a challenge because different specialists annotate images with discrepancies when given the same instructions. And, in some cases, these annotations are mutually exclusive, which results in a noisy dataset. Our contribution to this topic is the creation of a four-stage aggregation algorithm that makes it possible to train models on the clinical consensus, with a small tweak in favor of the most experienced and best-performing specialists. We call this method *clinical knowledge unification*, and it is a novel algorithm that consolidates the subjective estimations of a number of annotators, to serve as reliable ground truth for severity estimation, in the absence of a gold standard. The four stages of our algorithm are presented below (Figure 1):

- Step 1: Separate all annotations, that is, bounding boxes, by lesion type (abscess, nodule, and draining tunnel), and sort each cluster in decreasing order according to box area. This means that bigger boxes will be processed before the smaller ones. Every box is assigned one vote.
- Step 2: The biggest box of a lesion group is taken. Any other box that is inside or overlaps it more than a certain threshold is merged. This converts a set of N overlapping boxes into a single box with N votes, which is the bounding box of all the boxes involved (see boxes A and D in Figure 1, step 2). All the merged boxes are removed from the initial list and the process is repeated by picking the new biggest box. This is done until there are no more boxes to process. Any box that may escape this merging criterion during this iterative process is considered a valid box on its own and is also removed from the initial box list. This is done for every lesion group separately. The goal of this step is to remove redundant boxes. However, this may result in two undesired scenarios:
  - Two lesions of different types coexist in the same image region. This happens every time the annotators disagree. This case is exemplified by the bottom left boxes (D) in Figure 1, step 2.
  - Independent or non-merged boxes that have only been annotated by one person. The bounding boxes on the bottom right of Figure 1, step 2 (B and C) are examples of this situation, where B comes from the senior dermatologist and C from one of the less experienced dermatologists
- Step 3: The second undesired scenario from the previous step is corrected by removing any independent box that comes from any person other than the senior dermatologist. The motivation behind this is that, if we only kept overlapping boxes, we may be leaving behind some relevant lesions only spotted by the more experienced dermatologist. Boxes labeled as B in Figure 1 are examples of such a scenario.
- Step 4: Every remaining box is compared to each other. If two boxes with different labels overlap, the one with fewer votes is discarded (Figure 1, box D). This is called majority voting. In case two overlapping boxes have the same number of votes, we keep that with the most severe lesion type (nodule < abscess < draining tunnel).

The two criteria (Equation 1) used for merging boxes in Step 1 are Intersection over Union (IoU) and the ratio between the intersection

TABLE 2 *Legit.Health-HS-IHS4* dataset, as seen by the six annotators: Minimum and maximum lesion counts in an image, the average number of lesions detected in an image, and total number of lesions annotated.

	Abscesses				Nodules				Draining tunnels			
Annotator	Min	Mean	Max	Total	Min	Mean	Max	Total	Min	Mean	Max	Total
1	0	0.42	4	93	0	0.83	8	183	0	0.65	5	144
2	0	0.47	5	103	0	0.98	5	217	0	1.34	11	297
3	0	0.37	4	82	0	0.95	5	210	0	1.44	10	319
4	0	1.1	9	244	0	1.12	8	247	0	2.9	16	640
5	0	0.2	4	45	0	1.07	11	236	0	1.83	19	405
6	0	0.29	3	64	0	1.16	8	256	0	1.52	10	337

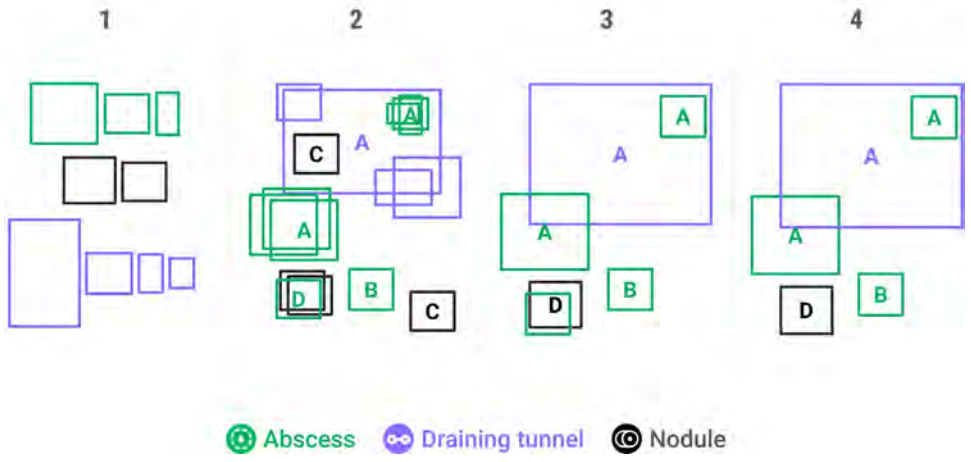

FIGURE 1 An example of our box aggregation algorithm. Each color corresponds to a lesion type (nodules, abscesses, and draining tunnels are black, green, and purple, respectively). Step 1 corresponds to box sorting. In steps 2 and 3, all boxes will be merged when possible (A), and others (C) will be removed unless they were annotated by the senior dermatologist (B). Majority voting (step 4) is finally applied to clean areas of the image with competing bounding boxes (D).

and the smaller box, which we called overlapping ratio (OR). The reason behind using the second criterion is that IoU is not enough when two overlapping boxes have very different sizes as it would be almost zero (see Figure 2). On the contrary, if box B is much smaller than box A, and most of it is inside it, OR would be close to one. Another reason to be more permissive by merging boxes with IoU and OR is due to the unacceptable results obtained with IoU alone: due to the variability between observers, the merged ground truth ended up with too many lesions, leading to extremely and unrealistically high IHS4 scores (see Figure 3). In summary, we avoid labels that are too specific while making sure that all relevant lesions are present in the merged annotations.

$$\text{IoU}(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad OR = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{1}$$

Majority voting, however, only focuses on IoU to find competing boxes. The reason behind using IoU as the only criterion in the third step is that there might be strong agreements between annotators in a lesion that is inside a bigger box of a different lesion with a similar agreement. Such a scenario might happen with small lesions such as nodules that are on top of other bigger and less superficial lesions such as draining tunnels.

2.3 | AIHS4

IHS4 is a validated tool to dynamically assess HS severity and can be used both in real-life clinical practice and clinical trials.[6] The resulting IHS4 score (Equation 2) is calculated by a weighted sum of the number of nodules (n), abscesses (a), and draining tunnels (t). A total score of 3 or less signifies mild, 4 to 10 signifies moderate and 11 or higher signifies severe disease.

$$\text{IHS4}(n,a,t) = n + 2a + 4t \tag{2}$$

In this study, and for the first time, we introduce the use of a deep learning model that automatically counts nodules, abscesses and draining tunnels, just by looking at a clinical image. We named this new tool AIHS4, which aims to overcome the inherent limitations of the IHS4 as a manual scoring system.

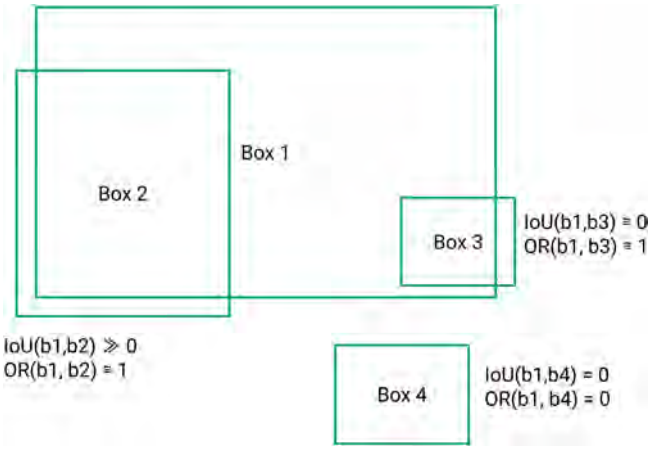

FIGURE 2 A dummy example to understand why it is necessary to use overlapping ratio (OR) in addition to Intersection Over Union (IoU). Box 1 is taken as the reference box, and all the other boxes are compared to it by means of IoU and OR. If we did not use OR, IoU would miss box 3 in the merging process since its value is almost zero.

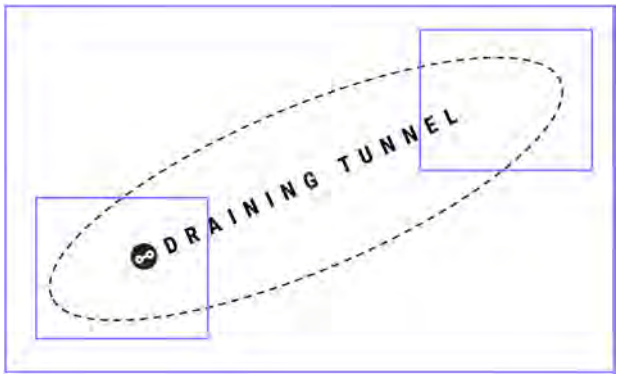

FIGURE 3 An example of how using Intersection over Union (IoU) alone would lead to exaggerated IHS4 scores. This figure conceptualizes a real case we found when reviewing the dataset: one dermatologist marked the exits of a single tunnel as separate lesions (small boxes), whereas the senior dermatologist correctly marked both as a single lesion (big box). However, the IoU between the small boxes and the big one is low, which would result in three separate lesions instead of one. By adding the overlapping ratio (OR) criterion, we overcome this problem: the smaller boxes are treated as redundant and removed, keeping the bigger box only.

2.3.1 | Deep learning model

Calculating IHS4 involves counting the number of three different types of lesions. Despite the regression-like nature of the problem (predicting an actual IHS4 score is what we need), we framed it as an object detection task, calling our lesion counting neural network *Legit.Health-IHS4Net*.

2.3.2 | Legit.Health-IHS4Net

Object detection is the task of detecting instances of objects of a certain class within an image. The state-of-the-art methods can be categorized into two main types: one-stage methods and two stage-methods. Single-stage methods prioritize inference speed, and example models include YOLO,[11] SSD[12] and RetinaNet.[13] Two-stage methods prioritize detection accuracy, and example models include Faster R-CNN,[14] Mask R-CNN[15] and Cascade R-CNN.[16]

We chose the YOLO architecture for this study. We used YOLOv5, which is an open-source implementation of YOLO that is extensively used by the machine learning community. It has a variety of architectures with an increasing number of parameters: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. *Legit.Health-IHS4Net* is a YOLOv5 trained for the detection of three types of lesions: nodules, abscesses, and draining tunnels. Apart from the total number of classes to be detected, no major changes were made to the YOLOv5 implementation.

The output of the Legit.Health-IHS4Net model (θ), given an input image k_i, is a set of N bounding boxes, each of them defined by six attributes (o_j, c_j, x_j, y_j,w_j,h_j): an *objectness* score o_j, that is, a probability value that indicates how likely is that box to contain an actual lesion; its location and size (x_j, y_j,w_j,h_j); and its corresponding predicted class c_j. By applying a method called non maximum suppression (NMS), we remove all the predictions with an objectness score below a certain threshold as well as redundant overlapping boxes (i.e., more than one box per actual lesion) by setting the appropriate IoU threshold.

Once the output is processed, it is easy to count the total number of abscesses, draining tunnels and nodules detected in an image ($n_i^\theta, a_i^\theta, t_i^\theta$) and calculate the corresponding IHS4 score predicted by model, $\hat{y}_i^\theta$ (Equation 3).

$$\begin{aligned} \theta(k_i) &= (o_j, c_j, x_j, y_j, w_j, h_j)\ j = 0 \ldots N \\ NMS(\theta(k_i)) &= \left\{ n_i^\theta, a_i^\theta, t_i^\theta \right\} \\ \hat{y}_i^\theta &= IHS4\left(n_i^\theta, a_i^\theta, t_i^\theta \right) = n_i^\theta + 2a_i^\theta + 4t_i^\theta \end{aligned} \tag{3}$$

2.3.3 | Experimental setup

We fine-tuned all four pre-trained YOLOv5 architectures (YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x) on the *Legit.Health-HS-IHS4* dataset. Each model was trained following a 6-fold cross-validation strategy, and each experiment was run for 300 epochs, with a batch size of 16, and an input image size of 640 × 640. Due to the reduced dataset size, we decided to apply some data augmentation techniques to make the most of the data available. We applied random horizontal and vertical flipping ($p_h = 0.5$, $p_v = 0.25$) and rotation ($\pm$ 25 degrees). The rest of the hyperparameters and data augmentation settings were set to default. These experiments were entirely run on a single NVIDIA Tesla V100 (32GB) graphics processing unit (GPU).

2.4 | Metrics

We chose mean absolute error (MAE) between the model's IHS4 prediction $\hat{y}_i^{\theta}$ and the ground truth y_i (Equation 4) as the main evaluation metric for IHS4 assessment. The reason behind using this regression metric is that the main goal is to calculate IHS4, rather than simply spotting HS lesions. Each dermatologist was also compared to the generated ground truth y_i by means of MAE (Equation 4). To get a better clinical understanding of both dermatologists and model performance in terms of MAE, we also computed this metric in separate severity groups s (Equation 4). In order to make consistent comparisons, we kept the same severity criterion through all experiments: every image of the dataset (k_i) was assigned a severity according to the IHS4 score of the senior dermatologist d_3 (Equation 5).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i^{\theta} - y_i\right| \quad \text{MAE}_d^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\left|\hat{y}_i^{d} - y_i\right| \tag{4}$$

$$s(k_i) = \begin{cases} \text{mild} & \text{IHS4}^{d_3}(k_i) \leq 3 \\ \text{moderate} & 3 < \text{IHS4}^{d_3}(k_i) \leq 10 \\ \text{seuere} & \text{IHS4}^{d_3}(k_i) > 10 \end{cases} \tag{5}$$

$$CV_{IHS4}(k_i) = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma_{IHS4}(k_i)}{\mu_{IHS4}(k_i)} \quad ACV_{IHS4} = \frac{1}{N}\sum_{i=1}^{N} CV_{IHS4}(k_i) \tag{6}$$

To get a better sense of annotation variability in an image k_i, we use the coefficient of variation or CV (Equation 6). The ratio between standard deviation and mean helps us understand scattering in IHS4 assessment. In some analyses, we also used the average coefficient of variation (ACV) to aggregate all coefficients of variation (see Figure 4). Apart from the regression metrics, we also worked with some object detection metrics, such as precision (P) and recall (R).

2.5 | CADx system

With the purpose of making the AIHS4 tool accessible to healthcare professionals, both in clinical trials and in routine evaluations, we created a fully integrated CADx system. The CADx system consists of a web application connected to the *Legit.Health-IHS4Net* model via an application programming interface (API) that calculates the patient-reported AIHS4 just by looking at clinical images that patients themselves uploaded from their phones.

To understand how the CADx system works, it is better understood as a three-stage process: uploading the images of the affected areas, processing the images, and reporting the AIHS4.

In the first stage, the patient uploads an image (I_i) of the affected area or areas to the system. In the second stage, the *Legit.Health-IHS4Net* model (θ) processes the image and automatically calculates the severity of HS by detecting the lesions and counting the number of each class, as they appear in the image ($n_i^{\theta}, a_i^{\theta}, t_i^{\theta}$). Finally, the model outputs a variety of clinical endpoints that are displayed through a user-friendly report. Said report contains the image with the estimated lesion surface and a chart with the evolution of AIHS4 across different instances of time, among other contextual information. An overview of the CADx system's report is depicted in Figures 5 and 6. If the user uploads images for different parts of the body, the CADx system calculates the global AIHS4 score (Equation 7) by using the following formula that combines N images of the whole body:

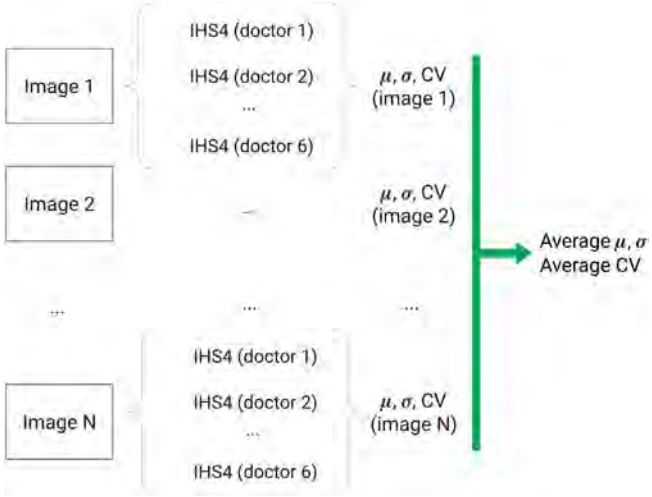

FIGURE 4 Visual explanation of how to obtain the average coefficient of variation (ACV). Every image has its own CV according to its corresponding IHS4 scores. ACV is the mean of all coefficients, and the average mean and standard deviation can also be computed.

$$\begin{aligned} AIHS4(I_i;\theta) &= IHS4\left(n_i^{\theta}, a_i^{\theta}, t_i^{\theta}\right) = n_i^{\theta} + 2a_i^{\theta} + 4t_i^{\theta} \\ AIHS4 &= \sum_{i}^{N} AIHS4(I_i;\theta) \end{aligned} \tag{7}$$

3 | RESULTS

3.1 | Annotation

3.1.1 | Individual clinical assessment

In this section we present the aforementioned overall inter-observer variability in terms of CV and MAE. By computing the coefficient of variation of every image, we find images with *soft disagreement* (left side) and others that yield *strong disagreement* between observers (right side), as seen in Figure 7. Table 3 describes annotation variability separated into severity groups according to IHS4 criteria (Equation 5): mild, moderate, and severe. This table shows more disagreement (higher coefficient of variation) between dermatologists when labeling mild cases.

Table 4 presents another perspective on the same variability problem. We compared every dermatologist to each other in all six

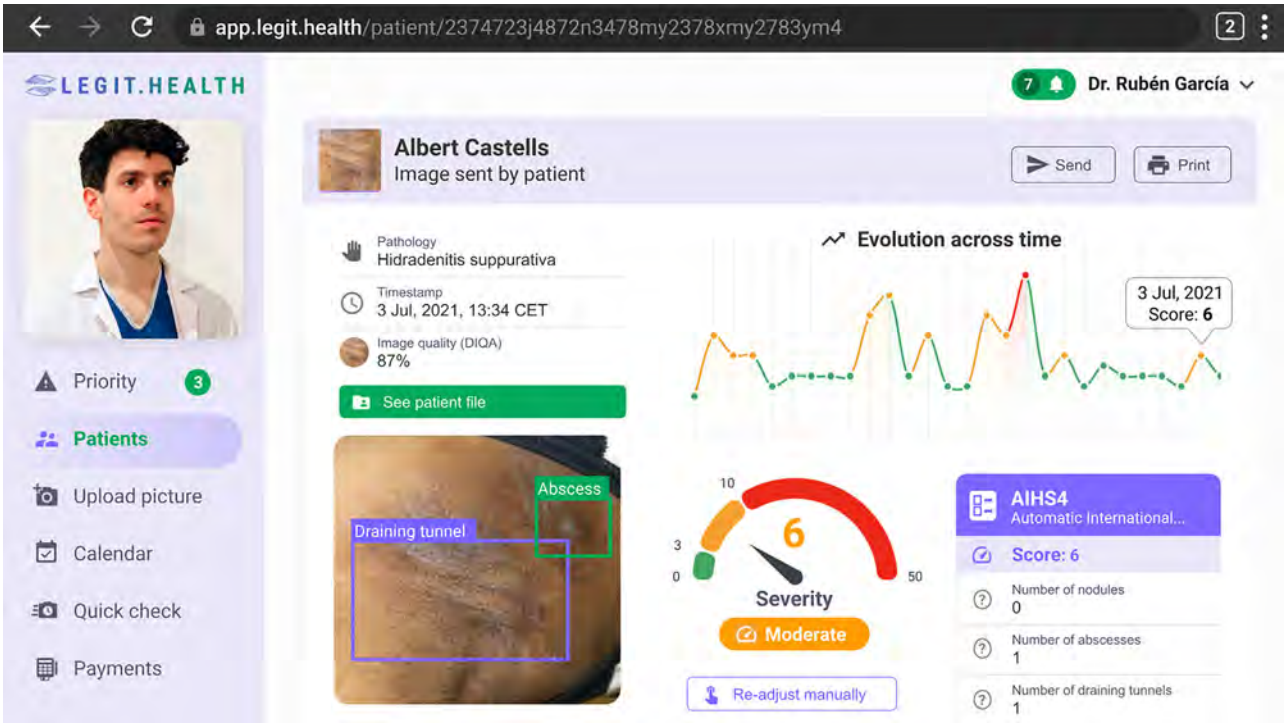

FIGURE 5 Caption of the full report on a CADx system. The full report generated by the model shows the evolution across time of the pathology in the framework of the AIHS4. The report highlights the lesions detected by the model, as they are used to automatically calculate IHS4, thus allowing the physician to supervise the model's performance. Previous IHS4 scores are also presented, which makes it possible to assess the current HS therapeutic strategy. It also shows the image of the latest case uploaded, with the number of lesions detected highlighted by their corresponding bounding boxes, and some additional information such as image quality.

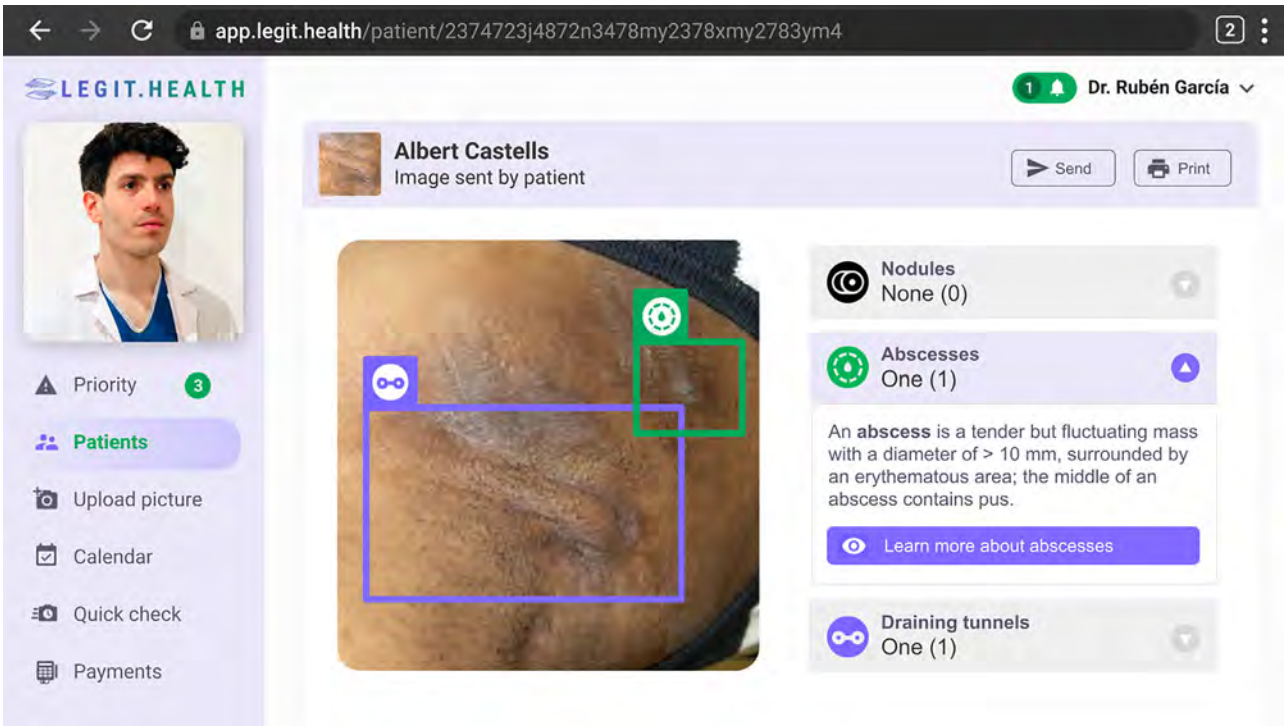

FIGURE 6 Caption of the lesion-detail report in the CADx system. The lesion-detail report highlights the type of lesions detected by the model, thus allowing the physician to supervise the model's performance in a quick and visual way. The lesions are highlighted by their corresponding bounding boxes and contain additional information about the lesions themselves.

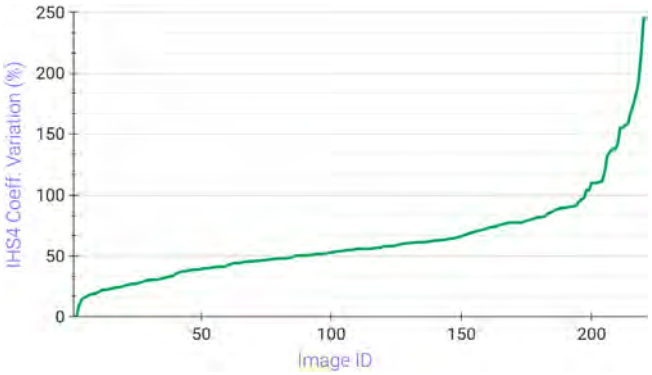

FIGURE 7 Coefficient of variation of IHS4 of every image of the dataset. Images are sorted by increasing CV, which reveals images with soft disagreement (left tail) and very strong disagreement (right tail).

TABLE 3 Variability of IHS4 scoring by severity level. Average CVs suggest strong disagreement between dermatologists, specially when labeling mild lesions. Despite still being high, there is less disagreement when labeling moderate and severe lesions. A visual explanation of the average coefficient of variation is shown in Figure 4.

Severity	IHS4 average coefficient of variation (%)
Mild (IHS4 ≤ 3)	100.23
Moderate (3 < IHS4 ≤ 10)	53.73
Severe (IHS4 > 10)	51.82

TABLE 4 One-versus-all analysis: For every image in each severity group, IHS4 scores of each dermatologist were compared to the others. Results were aggregated by averaging the mean absolute error (MAE) of all images. The table shows the results of doing this process in all validation splits and averaging the results.

	MAE (vs. every other)		
Dermatologist	Mild	Moderate	Severe
1	2.43	4.89	7.45
2	2.30	3.53	6.20
3	2.27	3.52	6.67
4	2.96	4.82	11.27
5	2.65	4.00	7.95
6	3.25	3.84	8.03
Group 1	2.64	4.10	7.93
Group 2	2.72	4.22	8.18

validation splits to observe the differences in IHS4 scoring. The second and third dermatologists achieved the lowest (best) MAEs in mild, moderate, and severe lesions. Finally, Table 2 was already presented in a previous section and summarises the number and type of lesions detected by each specialist.

TABLE 5 Mean absolute error between each dermatologist and merged ground truth, grouped by IHS4 severity. For the sake of better comparisons, severity groups were created based on the senior dermatologist's choices as in Table 3. This table shows the average of the six validation splits. The senior dermatologist achieved the strongest agreement with the ground truth in moderate and severe cases, and the second best mean absolute error (MAE) in mild cases.

	MAE (vs. ground truth)		
Dermatologist	Mild	Moderate	Severe
1	2.64	4.67	5.78
2	1.47	2.67	4.04
3	2.47	1.69	2.29
4	2.43	3.65	7.74
5	1.64	2.61	5.22
6	1.60	2.74	4.21
Group 1	2.04	3.01	4.88
Group 2	1.90	3.16	5.37

TABLE 6 Model performance in terms of mean absolute error (MAE) (six-fold cross-validation). Mean absolute error is split into the same three severity groups (mild, moderate and severe) of Tables 4 and 5. The sum of MAEs (mild, moderate, and severe) was used as the criterion to pick the best configuration of each model. IoU threshold was kept the same in all experiments (0.5). Every threshold setting was applied to every fold.

Model	Threshold (confidence@IoU)	MAE (mild)	MAE (moderate)	MAE (severe)	Sum
yolov5s	0.10@0.50	2.49	3.62	5.35	11.46
yolov5m	0.20@0.50	2.21	3.22	6.17	11.60
yolov5l	0.20@0.50	2.8	3.72	5.52	12.04
yolov5x	0.20@0.50	2.16	3.37	5.26	10.79

3.1.2 | Combined clinical assessment

After running the merging algorithm described earlier in this work, we compared each dermatologist to the obtained consensus by means of Mean Absolute Error. The third annotator, that is, the senior dermatologist, shows the strongest agreement with the generated ground truth (Table 5) in moderate and severe cases: this is reasonable and expected since the algorithm gave preference to his annotations. However, we didn't observe a decrease in MAE for mild lesions.

3.2 | Legit.Health-IHS4Net

Model performance in terms of MAE is summarised in Tables 6 and 8. On average (i.e., all six validation splits), using the threshold settings presented in the table yielded the best results for each architecture. Grouping by lesion severity helps to understand the real performance of the models: for example, the clinical relevance of a 3-point MAE in

TABLE 7 Object detection metrics: Precision (P) and recall (R). The values in the table are means ± standard deviations. All models present similar performances with six-fold cross-validation.

Model	Precision (P)	Recall (R)
yolov5s	0.44 ± 0.09	0.40 ± 0.12
yolov5m	0.42 ± 0.08	0.39 ± 0.13
yolov5l	0.45 ± 0.08	0.41 ± 0.11
yolov5x	0.46 ± 0.10	0.39 ± 0.07

a severe case is much different from that of a mild or moderate case with the same MAE. We explored different *objectness* thresholds t_o while keeping the same Intersection Over Union (IoU) threshold t_{IoU}. The criterion for selecting the best model was the sum of all MAEs (Equation 8): in terms of this metric, YOLOv5s and YOLOv5m yielded the bests results on this task with an overall MAE of 12.74 and 12.65, respectively.

$$\begin{aligned}\mathrm{MAE}(\theta; t_o, t_{IoU}) = {} & \mathrm{MAE}_{\mathrm{mild}}(\theta; t_o, t_{IoU}) \\ & + \mathrm{MAE}_{\mathrm{moderate}}(\theta; t_o, t_{IoU}) + \mathrm{MAE}_{\mathrm{severe}}(\theta; t_o, t_{IoU})\end{aligned} \tag{8}$$

Other object detection metrics, such as precision and recall, are detailed in Table 7. In terms of precision and recall, YOLOv5s yielded similar results ($p = 0.44, R = 0.40$) to YOLOv5l ($p = 0.45$, $R = 0.41$). However, if we consider how model size affects performance, YOLOv5l becomes the preferred model. This, together with the presented MAEs, presents YOLOV5s as the best model of all for this task.

4 | DISCUSSION

In this work, we have faced the challenges of (1) gathering a clinical board to generate an object detection image database and (2) training a family of deep learning models to solve the task of automatic IHS4 assessment.

In the first place, we observed a remarkably high inter-observer variability. For example, results presented in Table 3 may be due to the less experienced dermatologists over-detecting lesions in a mild case (according to senior criteria), or the senior dermatologist marking fewer or no lesions in an image that everyone else considered to be of a more severe case (i.e., under-detection). In other words, images labeled as mild by the senior dermatologist may be labeled as moderate or severe by the rest of the board, depending on which and how many lesions they found.

The annotation results presented in the previous section revealed the necessity of a merging algorithm to aggregate all experts' opinions and find a consensus. As stated before, we preferred to define an algorithm that generalizes to all annotators' opinions instead of one that just roughly merges all annotations by means of IoU. Naively merging boxes entirely based on IoU is particularly undesirable with draining

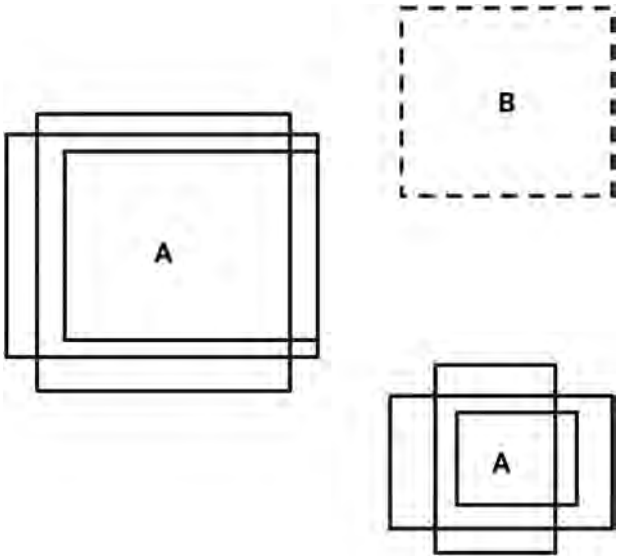

FIGURE 8 An example of the effect of merging labels on IHS4 mean absolute error. In this case, the image would be labeled as mild because the senior dermatologist (B) only found one lesion. However, the other dermatologists (A) found more lesions in similar places. Assuming all lesions are nodules, the annotation from the senior dermatologist would yield an IHS4 of 1, whereas the ground truth resulting from using our unification method would be 3. This results in an absolute error of 2, and more cases like this would eventually result in a higher mean absolute error (MAE).

tunnels, as every lesion of this category results in a 4-point increment in the final IHS4 score. This means that we had to be very careful when defining the unification method. Figure 3 conceptualizes an example of a real case in which we dealt with extreme IHS4 scores due to draining tunnels being labeled differently by the dermatologists. By following our method, that is, merging labels into bigger bounding boxes and discarding others in a more selective manner, we avoided overestimated IHS4 scores while staying confident that any relevant lesion would be present in an image. We observed that the senior dermatologist yielded a slightly higher MAE in mild cases (Table 5). The reason behind this could be the majority voting applied in the last step of our merging algorithm: if a lesion is detected by three or more annotators, it will be added to the final ground truth, even if it is not detected by the senior dermatologist. Figure 8 presents an example of this scenario.

In terms of model performance, after validating the model we observed that, on average, it was behind that of any of the dermatologists when assessing mild cases. We strongly believe that such a problem would be fixed by using a larger dataset. However, the results are compelling: some of the models got closer to the ground truth than some specialists (see Table 8).

Another reason for the current model performance is the difficulty in labeling the *Legit.Health-HS-IHS4* dataset in a retrospective manner: IHS4 assessment involves lesions that require palpation and close examination, which is not possible in remote care. This might explain why it was difficult to achieve consistent labels among annotators. Due to not being able to assess the patient in person, dermatologists are more prone to make mistakes or disagree with each other, as it is

TABLE 8 Comparing dermatologists and model performance to ground truth annotations (average of six validation splits).

	Mean absolute error (MAE) (vs. ground truth)		
Method	Mild	Moderate	Severe
Dermatologist 1	2.64	4.67	5.78
Dermatologist 2	1.47	2.67	4.04
Dermatologist 3	2.47	1.69	2.29
Dermatologist 4	2.43	3.65	7.74
Dermatologist 5	1.64	2.61	5.22
Dermatologist 6	1.60	2.74	4.21
Group 1	2.04	3.01	4.88
Group 2	1.90	3.16	5.37
yolov5s	2.49	3.62	5.35
yolov5m	2.21	3.22	6.17
yolov5l	2.8	3.72	5.52
yolov5x	2.16	3.37	5.26

challenging to estimate real lesion size and appearance from a single image. Using a dataset with images labeled at the exact time of patient exploration would avoid these inconsistencies and lead to better results. Due to the subcutaneous nature of HS, we believe another source of improvement could be expanding conventional skin surface image datasets with ultrasound images.[17] Providing the model with such clinical and ultrasound image pairs could also improve detection performance, given the appropriate sample size, or at least make annotators more confident when labeling a clinical image.

Apart from the method used for gathering the dataset, we believe another reason behind these results is the nature of the images: there are noticeable differences in terms of picture quality. According to the results of our annotation survey 8, it is possible that specialists found this task difficult due to this nonstandard picture quality. This, apart from the evident differences in clinical expertise, could contribute to the high variability. Gathering a larger dataset with more consistent picture quality and acquisition settings (lighting, distance to object, imaging device, etc) could improve both annotation agreement and model performance.

Clinical disagreement, that is, annotation variability, supports our idea of using AI in the treatment of HS. By merging clinical knowledge and training AI models on the clinical consensus of larger clinical boards, it could be possible to stay ahead of the disease: our model could spot emerging severe lesions when a physician is only observing mild or moderate signs. Extending the use of such an AI tool would lead to a higher number of specialists staging their patients, resulting in a better-documented disease and the creation of strongly validated treatment guides.

5 | CONCLUSION

In this work we have presented the AIHS4, the first AI-based model that automatically fills in the IHS4 scoring system by looking at clinical images. The main advances of this algorithm are reducing the time spent by physicians in filling in the manual severity scoring system and standardizing HS assessment with reduced inter-observer variability. Automated HS assessment is done by a state-of-the-art object detector, YOLOv5, that was trained on the *LegitHealth-HS-IHS4* dataset containing HS images with their corresponding IHS4 scores.

Despite the lack of a large image dataset and the limited size of the clinical annotation team (with different years of experience in assessing HS), we consider this work to be a successful proof of concept with promising results. We were able to overcome clinical assessment variability by developing a merging algorithm that fuses all experts' annotations to create a consensus. This will become an essential tool when dealing with a bigger dataset annotated by more experts.

In conclusion, we believe that our model has the potential to reduce costs in dermatology by saving time, whilst improving documentation of the evolution of HS.

6 | LIMITATIONS

As discussed before, the presented model would benefit from a much larger image dataset with more consistent picture quality. In future works, our goal is to reach a dataset size one or two orders of magnitude larger, making sure that we collect as many cases of every skin tone as possible to overcome any bias related to this factor. Having the data annotated by a much larger clinical board would boost performance by means of a stronger clinical consensus generated with our merging algorithm. Regarding the clinical knowledge unification method, it is trivial to modify the algorithm so that it does not favor any specific annotator and treats all of them equally, in case the annotation board is larger enough. Data quality might also be improved by collecting the images at the time of the physical examination, allowing for palpation and ultrasound imaging to confirm the presence of lesions in the image.

An additional limitation to take into account is that patients affected in many areas might need to take more than one picture, slowing down the process of calculating the AISH4. Getting the best settings for taking pictures (e.g., lighting, background, lesion focus) can also impact the time spent in AISH4 follow-up. Moreover, if a photo of some affected area is not taken (e.g., due to the discomfort of taking and uploading a picture of genitalia or the patient's inability to take a picture of a difficult body part without directly seeing a preview in their imaging device), the CADx system's output won't be the actual AISH4.

However, the process of taking the images does not necessarily have to be done by the physician himself. This keeps the potential of our model to reduce costs in teledermatology by saving time and generating a valuable record of the evolution of the disease, which could be interesting for some scenarios such as pharmaceutical clinical trials.

ACKNOWLEDGMENTS

The authors thank IBM and Amazon Web Services (AWS) for providing the computing infrastructure for the deep learning experiments, and

BioCruces Bizkaia Health Research Institute for the academic support. This project has been funded by the Department of Economic Development and Infrastructures of the Basque Government (HAZITEK Program) and the European Regional Development Fund (ERDF).

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

FUNDING INFORMATION

This project has been funded by the Department of Economic Development and Infrastructures of the Basque Government (HAZITEK Program) and the European Regional Development Fund (ERDF).

DATA AVAILABILITY STATEMENT

The images used in this work come from Dermnet NZ (https://dermnetnz.org) and DermQuest. The DermQuest dataset is available as part of another widely used skin image dataset, SD-198.[18] The hidradenitis suppurativa images of this dataset were selected for this work.

ORCID

Ignacio Hernández Montilla https://orcid.org/0000-0003-0356-6619
Alfonso Medela https://orcid.org/0000-0001-5859-5439
Taig Mac Carthy https://orcid.org/0000-0001-5583-5273
Andy Aguilar https://orcid.org/0000-0003-0618-6179
Pedro Gómez Tejerina https://orcid.org/0000-0002-1379-6541
Alejandro Vilas Sueiro https://orcid.org/0000-0002-2681-5254
Ana María González Pérez https://orcid.org/0000-0002-4702-2659
Laura Vergara de la Campa https://orcid.org/0000-0001-8646-3636
Loreto Luna Bastante https://orcid.org/0000-0003-2712-8752
Rubén García Castro https://orcid.org/0000-0001-8299-1706
Fernando Alfageme Roldán https://orcid.org/0000-0002-0962-9783

REFERENCES

1. Zouboulis C, del Marmol V, Mrowietz U, Prens EP, Tzellos T, Jemec GBE. Hidradenitis suppurativa/acne inversa: criteria for diagnosis, severity assessment, classification and disease evaluation. *Dermatology*. 2015;231:184–190. https://doi.org/10.1159/000431175
2. Paus LR, Kurzen H, Kurokawa I, et al. What causes hidradenitis suppurativa? *Exp dermatology*. 2008;17:455–456. https://doi.org/10.1111/j.1600-0625.2008.00712_1.x
3. Fimmel S, Zouboulis CC. Comorbidities of hidradenitis suppurativa (acne inversa). *Dermato-Endocrinology*. 2010;2:9–16. https://doi.org/10.4161/derm.2.1.12490
4. Alikhan A, Lynch PJ, Eisen DB. Hidradenitis suppurativa: a comprehensive review. *J Am Acad Dermatol*. 2009;60:539–561.
5. Ingram JR, Hadjieconomou S, Piguet V. Development of core outcome sets in hidradenitis suppurativa: systematic review of outcome measure instruments to inform the process. *Br J Dermatol*. 2016;175:263–272.
6. Zouboulis C, Tzellos T, Kyrgidis A, et al. Development and validation of ihs4, a novel dynamic scoring system to assess hidradenitis suppurativa/acne inversa severity. *Br J dermatol*. 2017;177:1401-1409. https://doi.org/10.1111/bjd.15748
7. Pala P, Bergler-Czop BS, Gwiżdż, JM. Teledermatology: idea, benefits and risks of modern age–a systematic review based on melanoma. *Adv Dermatol Allergol Dermatol i Alergologii*. 2020;37:159.
8. De A, Sarda A, Gupta S, Das S. Use of artificial intelligence in dermatology. *Indian J Dermatol*. 2020;65:352.
9. Welikala RA, Remagnino P, Lim JH, et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access*. 2020;8:132677–132693.
10. Khudorozhkov R, Koriagin A, Kozhevin A. Clearing noisy annotations for computed tomography imaging. Paper presented at: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS); October 15-18, 2018; Valencia, Spain.
11. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016 (pp. 779–788).
12. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. *Lect Notes Comput Sci*. 2016;21–37. https://doi.org/10.1007/978-3-319-46448-0_2
13. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* 2017 (pp. 2980–2988).
14. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*. 2015;28.
15. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer* vision 2017 (pp. 2961–2969).
16. Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 2018 (pp. 6154–6162).
17. Roldán FA. Ultrasound skin imaging. *Actas Dermo-Sifiliográficas English Ed*. 2014;105:891–899.
18. Sun X, Yang J, Sun M, Wang K. A benchmark for automatic visual classification of clinical skin disease images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands*, October 11-14, 2016, Proceedings, Part VI 14 2016 (pp. 206–222). Springer International Publishing.

How to cite this article: 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. https://doi.org/10.1111/srt.13357