AcneAI: A new acne severity assessment method using digital images and deep learning *

Léa Gazeau[1,2], Hang Nguyen[1,2], Zung Nguyen[1,2], Mariia Lebedeva[1,2], Thanh Nguyen[1,2], Tat-Dat To[3], Jimmy Le Digabel[4], Jérome Filiol[4], Gwendal Josse[4], Clifford Perlis[2,5], and Jonathan Wolfe[2,6]

[1] Torus AI, 12 Av. de l'Europe, Ramonville-Saint-Agne, France
[2] BelleTorus Corp., 245 First Street Riverview II, 18th Floor Cambridge, MA, USA
[3] IMJ-PRG, Sorbonne Université, Paris, France
[4] Centre de Recherche sur la Peau, Pierre Fabre Dermo-Cosmétique, France
[5] Keystone Dermatology Partners, King of Prussia, PA, USA
[6] Jefferson-Einstein Montgomery Medical Center, East Norriton, PA, USA

Abstract. In this paper we present a new AcneAI system that automatically analyses facial acne images in a precise way, detecting and scoring every single acne lesion within an image. Its workflow consists of three main steps: 1) segmentation of all acne and acne-like lesions, 2) scoring of each acne lesion, 3) combining individual acne lesion scores into an overall acne severity score for the whole image, that ranges from 0 to 100. Our clinical tests on the Acne04 dataset shows that AcneAI has an Intraclass Correlation Coefficient (ICC) score of 0.8 in severity classification. We obtained an area under the curve (AUC) of 0.88 in detecting inflammatory lesions in a clinical dataset obtained from a multi-centric clinical trial.

Keywords: Computer vision, Deep learning, Acne segmentation, Acne severity assessment, Acne scoring, Acne classification

1 Introduction

Acne vulgaris (acne for short) consistently represents one of the top three most prevalent skin conditions in the world's population [6]. According to some studies [17,12], at every moment about 20% of the world's population has active acne. About 85% of young adults aged 12 to 25 are affected by acne [2]. In 2022, we estimate that the total market of acne medication exceeded 9.9 billion in the U.S [9].

* This research is supported by Torus AI (https://torus.ai) and BelleTorus Corporation (https://belle.ai). H. Nguyen, Z. Nguyen, L. Gazeau, T. Nguyen and TD. Tô contributed in the technical part. M. Lebedeva, J. Digabel, J. Filiol and G. Josse contributed in the validation part. C. Perlis and J. Wolfe contributed in the medical part.
Corresponding author: gazeau.lea@torus.ai

There are two common methods for acne severity assessment [5,16,1]: counting and grading. Acne counting is a challenging task, because acne lesions are often small and can be easily missed or confused with other primary lesions. Grading is usually based on the comparison with "typical cases" of different grades, without a detailed counting. The grading method is easier to perform, but is also more subjective, and can be inconsistent among the clinicians.

There exist several research works dealing with acne severity and acne counting using deep learning [8,23,15,24,20,22,21] [1]. Some papers use the object detection approach which counts acne lesions and classifies them into (sub)types. In [8], the authors used Faster R_CNN [13] for acne detection and acne classification. They count the number of acne lesions of each type and use the LightGBM [10] method to measure global severity. The classification system divides lesions into only three different groups: comedones; papules-pustules; and nodules-cysts. In the grading step, they consider only the number but not the sizes of acne lesions, e.g., a very small nodule will have the same score as a very big nodule. In another study [21], the authors proposed a new acne severity assessment which considers area of each acne type. In their method, all acne of one type have the same score as they used multi-label segmentation. They did not provide acne counting. In [24], the authors developed deep learning models for severity grading of a full face (instead of just an image), by combining three images (left, right and front) together into a full face image. However, their approach does not provide acne counting. In [23], the authors used Label Distribution Learning and Fully Convolutional Network together to generate a distribution of the number of acne lesions and distribution of severity for each input image. Their method does not give the position for each acne lesion and does not classify acne into sub-types.

Our contribution

In AcneAI, all lesions, either acne or non-acne lesions that look like acne, are first segmented. This over-segmentation enables easier and fuller detection of acne lesions: data annotators who are not doctors or clinicians can be trained to annotate, and they will not miss many acne lesions. Doctors or clinicians are thus required only for the second step of this process –annotating the database by classifying the segmented lesions (on cropped images centered around these lesions) into non-acne and acne sub-types.

This approach has several major advantages compared to the approach of requiring clinicians to segment/count (and eventually classify) acne on original images: it saves time for the clinicians, provides more detailed segmentation (with precise shapes instead of just bounding boxes), and achieves less confusion in the database.

As shown in [18], the severity grading method, by which the doctors grade the severity of an acne image by comparing it to typical examples of varying severity grades, is subjective and leads to a high level of inconsistency (low

[1] Please see Table 2 in supplementary file for detail comparison

intra- and inter-clinician correlations). AcneAI solves this inconsistency problem by objectively counting every acne lesion, measuring its surface area (thanks to precise segmentation) and its local severity, and then combining all the individual severity scores together (by a non-linear mathematical formula) into a granular overall severity score that ranges from 0 (acne-free) to 100 (theoretical value for extremely severe acne). This acne severity score is sensitive to even small changes in the number or the severity of acne lesions, and is therefore useful for regular monitoring of the evolution of acne in a patient.

We manually checked and re-annotated the Acne04 dataset (Acne04 v2) and make it available for the research community. A live demo of AcneAI system is currently available at https://demo.belle.ai [1].

2 Detailed description of AcneAI

For an input image, we first run the segmentation model, which segments all acne and acne-like lesions. We then run the acne separation algorithm to identify the center and radius of every lesion, including those that share borders. Next, we crop the original image into many smaller images based on the list of centers and radii. This step outputs a series of acne-centered images to be scored using a regression model. In the last step, we compute acne severity for the input image based on the total number of acne lesions and the severity and area of each acne lesion. Table 1 in the supplementary file shows an overview of our training/ testing data.

2.1 Acne segmentation model

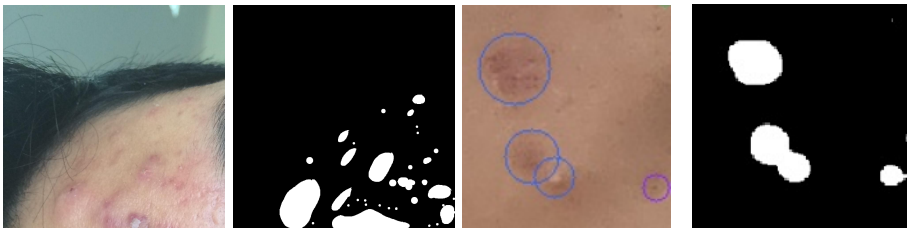
This model segments not only acne but also acne-like lesions, for example acne scars or moles. This is a small-object detection problem where the goal is detect all acne lesions, even the tiny ones. We use a dataset containing 901 images (AcneAI seg. data) which come from our collaborating doctors, and our partners (in Asia, the United States and Europe) taken in their clinics by smartphones and professional cameras. Since we segment all acne and acne-like objects, our process does not require doctors to do the job. The annotation uses a circle for small acne lesions (comedo, small papule/ pustule) and the exact shape for large lesions. The dataset was annotated by our trained annotators. Fig. 1a, 1b shows one example in our dataset.

We use a U-Net architecture [14] with EfficientNet B4 encoder [19] pre-trained on ImageNet [4], with an input size of 512×512 pixels. We trained the model using an Nvidia RTX 3090 GPU with a batch size of 7, training steps 315 for 200 epochs. We set up a learning rate schedule as $start\_lr * (0.99^{epoch})$.

2.2 Acne separation

A challenge in acne segmentation occurs when one lesion is right next to another, creating a risk that the two lesions will be seen as one. If they are not separated, it

[1] Please contact info@belle.ai to get a trial use account.



(a) Original image (b) Segmentation (c) Original image (d) Prediction

Fig. 1: Example of AcneAI segmentation data and an example of how two acne lesions right next to another are segmented as one.

will make the classification step incorrect, either because the two acne lesions are of different types, or because in seeing them as one the model will overestimate their area or size. We use geometrical methods and mathematical analysis to separate these lesions and classify them individually. That is we will shrink each contour of the predicted masks until we get separated regions (overlapping acne) or no region (only one acne). Fig. 1c, 1d show before and after separation.

We assume that the shape of each acne lesion is a circle. Therefore, during the annotation process, our annotators annotate acne lesions by a minimal circle that covers the full lesion. When we train the acne segmentation model with those data, the model will segment acne by a circle. If two lesions are close, it returns two overlapping circles.

2.3 Acne scoring and classification model

One challenge in classification is that doctors classify acne inconsistently, especially for borderline cases where a lesion could be classified, for example, as either a big comedo or a small papule, or as either a big papule or a small nodule. As a result, we design a scoring system for acne (depending on the size of the acne): not acne: 0; comedo: 1 to 2; papule/ pustule: 2 to 4; nodule/ cyst: 4 to 5.

We want the model to classify only the acne lesion that lies at the center. Therefore, we used the region of interest (ROI) technique to force the AI to focus on the center. That is, the input of the acne classification model consists of one RGB image and one ROI image. The outputs are the severity score, the presence of pus, and the presence of an acne scar. The first output has a ReLU activation function and the corresponding values are between 0 for a non-acne lesion and 5 for a cyst. The two other outputs have a Sigmoid activation function as the cases are binary: presence (1) or absence (0) of pus in the lesion, and presence (1) or absence (0) of an acne scar (see Figure 1 in supplementary file for the detail structure of the model). The model was designed based on the Xception structure [3] with a residual layer and multiplication step between the original image and the ROI image. This model is trained with 128×128 pixel images. Each cropped image may contain one or many acne lesions. Training batch size

is 40 and training steps is 500. The model was trained using a Nvidia RTX 3090 with a learning schedule as $start\_lr * 0.998^{(epoch/50)}$ and Adam optimizer [11].

The training data contains 12,682 images cropped from the annotated dataset. Each of these images was annotated by three different doctors using AllbyAI platform [1]. This annotation process is independent from the segmentation step. We take the average score of the doctors as our ground truth.

Loss function We constructed a loss function which takes into account all network's outputs, defined as follows:

$$L(Y, \hat{Y}) = MSE(y_1, \hat{y}_1) + BCE(y_3, \hat{y}_3) + F(Y, \hat{Y}) \tag{1}$$

Where $Y = (y_1, y_2, y_3)$ is the ground truth and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3)$ is the prediction, y_1, y_2, y_3 stand for each output (score, pus, and scar respectively). MSE is the mean square error, and BCE the binary cross-entropy. The function F is defined by the following formula with n being the batch size:

$$F(Y, \hat{Y}) = \frac{-1}{n} \sum_{i=1}^{n} (y_{2i} log(\widehat{y_{2i}}) + (1 - y_{2i}) log(1 - \widehat{y_{2i}})) \times (2 - min\,(2, |3 - y_{1i}|)) \tag{2}$$

F corresponds to the reduced sum over all samples of a batch, of the BCE for output pus, multiplied by a term that depend on the true score of the sample. This term allows us to only include this loss where the severity score of the lesion is high enough to consider the pus information.

2.4 Acne severity assessment

The overall acne severity score of a given image is computed as follows:

$$S = \frac{200}{\pi} arctan\left(20 \sum_{i=1}^{N} s_i \frac{a_i}{A}\right) \tag{3}$$

where N is total number of acne lesions, s_i is the score of acne i (from acne classification model), a_i is the area of acne i (from segmentation model) and A is the total area of detected skin [2], $\sum_{i=1}^{N} s_i \frac{a_i}{A}$ represents the *additive* severity score (which can go very high), and the nonlinear function arctan is used to transform this additive score into a score ranging from 0 to 100. The idea is that using a nonlinear function is similar to the use of nonlinear "utility" functions in economics and other fields.

[1] https://www.allby.ai

[2] We use a skin segmentation model to calculate this number.

3 Evaluation

3.1 Clinical dataset

A clinical trial was conducted on 55 volunteers recruited in eight investigational centers located in Europe. All subjects had oily skin and presented almost clear, mild or moderate acne severity. They were followed during one year and had to regularly take photos of their face using their smartphones. In this dataset, investigators only labeled inflammatory acne lesions. There are 768 images with 1,584 annotated lesions.

We used area under the curve (AUC) to measure the accuracy because this is a binary classification problem (inflammation vs. non inflammation lesion). All lesions have two scores, one by AcneAI (from 0 to 5) and another one by investigators (which is always 1). We obtained an **AUC of 0.88** when we compute the average true positive rate (TPR) and false positive rate (FPR) for each image and then aggregate them. Fig. 2 shows the overview of the prediction of AcneAI on every acne lesion.

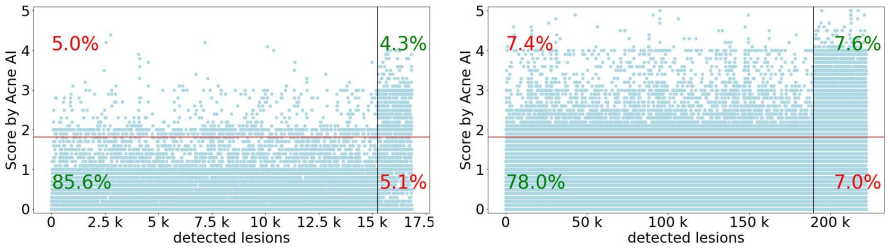


Fig. 2: Graphs show the score distribution of the AI on the clinical dataset (left) and Acne04 v2 (right). Each dot is one lesion. Every dot in the right side of the vertical black line is annotated lesion. The y-axis is the score by the AI and x-axis is the number of acne lesion. The horizontal red line is an example of threshold.

3.2 Acne04 dataset

We evaluated our acne severity assessment system on the Acne04 dataset [23] which is not used in our training/ validation. This dataset has bounding boxes for acne lesions but has no classification into sub-classes. It also provides overall severity grading (from 0 to 3) following the Hayashi criterion [7]. Looking at the Acne04 dataset, we found that: its bounding boxes created by experts are not lesion-centered and/or are too big with respect to the lesion; sometimes more than one lesion are in the same bounding box; and lesions are sometimes missed by their experts (See Fig. 3) For these reasons, we do not use the Acne04 dataset

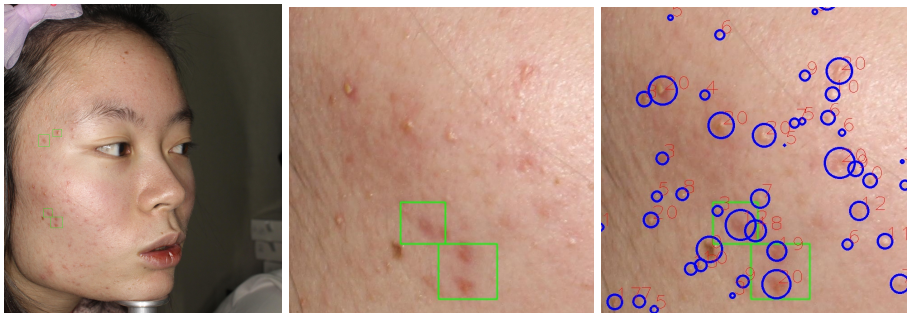


Fig. 3: Left is the original image with boxes, middle and right are zoom then crop versions. We can see many missing acne lesions on her cheek. Right image is our prediction with score (x10) for each acne.

to analyse our performance on acne counting. Since our system rarely misses any acne, the number of acne lesions for each image detected by our AI will be much higher than the label provided in the dataset. We evaluated our model on re-annotated Acne04 (see Data availability section below). We obtained an AUC of 0.79. With a threshold of 1.81 we obtain a true positive rate of 0.52 and false positive rate of 0.09 (Fig. 2 and Table 1).

Severity The Acne04 dataset has 4 levels of severity and AcneAI gives a continuous severity score for each image (value from 0 to 100, see Section 2.4). Fig. 4 shows the boxplots of our AI severity scores versus Acne04 severity levels, showing a high correlation between our AI severity scores and ones annotated by their experts. Another observation is that the dispersion of values is larger for higher levels. This is because the Acne04 level of severity mostly depends on the number of lesions, while our severity is calculated using the number, the size and the type of acne lesions. We convert our severity score (from 0 to 100) to 4 levels as in Acne04 as follows.

- We used an ROC curve to find the best threshold to separate the dataset into two groups: level 0-1 and level 2-3. The best threshold is 45.46.
- We repeated the process for those two groups and found that the best thresholds are: 17.86 for level 0 and level 1; and 78.86 for level 2 and level 3.

We obtain an **ICC score of 0.81** between our severity and Acne04 severity.

Data availability We manually checked and corrected wrongly labeled acne in Acne04 dataset. In this version, we use a circle which can be converted to box. The correction includes deleting low quality images, adding circles, and modifying circles such that each contains only one acne lesion. The re-annotated version of Acne04 is available at https://github.com/AIpourlapeau/acne04v2.

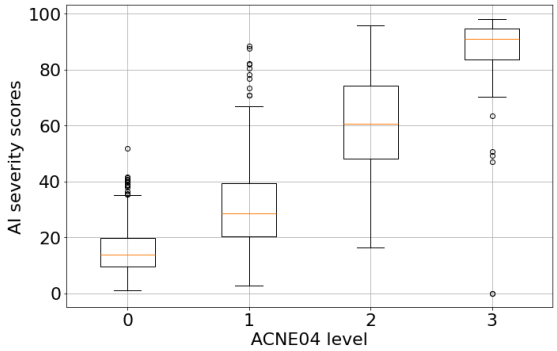


Fig. 4: Each box is one level by Acne04. Y-axis is severity score given by AcneAI. Non-overlapping boxes (lower, upper quartile) shows high correlation between severity by AcneAI and Acne04.

4 Comparison with other approaches

To compare our approach with another object detection approach, we trained several models using Ultralytics YOLOv8[1] (input size 1280x1280x3). We use acne separation (Section 2.2) to convert AcneAI seg. data to bounding box data which can be use to train YOLOv8. Table 1 shows the performance of YOLOv8 and AcneAI on different settings (Fig. 2 in supplementary file shows score distribution). Our results illustrate that AcneAI gives slightly better results. Moreover, the thresholds of YOLOv8 varies a lot compare to AcneAI.

5 Conclusions and future work

We propose an AcneAI system that can detect, classify and give global acne severity. Our system avoids the inconsistency of clinicians in detecting acne and achieves objective tracking of acne progression, which may drive better treatment outcomes for those suffering from chronic acne. There are some limitations that we will improve in the future: the prediction time is slow since the system contains two models; the scoring model scores each acne not the whole image at a time therefore the system is quite sensitive to the quality of the input image; in case of big nodules, Section 2.2 may give a big circle which can overlap with other small circles.

Acknowledgement. Authors would like to thank Paul Sherer and Ly Tran for a careful reading of the paper.

[1] https://github.com/ultralytics/ultralytics

Table 1: Performance of AcneAI and YOLOv8. First threshold is chosen by minimizing $|FP - FN|$, the second one is found by maximizing $(TP - FP)$. Here AP stands for Average precision, and corresponds to the number of correct predictions over the total number of predictions.

	Threshold	TPR	FPR	Precision	AP	AUC
Performance on Acne04 v2						
AcneAI	1.81	0.52	0.09	0.52	0.45	0.79
	2.01	0.18	0.02	0.64		
YOLOv8 (trained on AcneAI seg.)	0.18	0.54	0.14	0.54	0.57	0.72
	0.29	0.35	0.04	0.73		
Performance on clinical data						
AcneAI	1.81	0.46	0.06	0.46	0.41	0.88
	2.21	0.27	0.01	0.67		
YOLOv8 (trained AcneAI seg.)	0.25	0.33	0.05	0.34	0.30	0.88
	0.40	0.06	0.00	0.61		
YOLOv8 (trained on Acne04 v2)	0.16	0.32	0.04	0.32	0.26	0.85
	0.41	0.06	0.00	0.68		

Financial interests: Ms. Gazeau, Dr. Hang Nguyen, Dr. Zung Nguyen, Ms. Lebedeva, Dr. Thanh Nguyen, Dr. Perlis, and Dr. Wolfe are shareholders of Belle.ai.

References

1. Adityan, B., Kumari, R., Thappa, D.M.: Scoring systems in acne vulgaris. Indian J. Dermatol. Venereol. Leprol. **75**(3), 323–326 (May 2009)
2. Bhate, K., Williams, H.C.: Epidemiology of acne vulgaris. Br. J. Dermatol. **168**(3), 474–485 (Mar 2013)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807 (2017). https://doi.org/10.1109/CVPR.2017.195
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
5. Dréno, B., Thiboutot, D., Gollnick, H., Finlay, A.Y., Layton, A., Leyden, J.J., Leutenegger, E., Perez, M., on behalf of the Global Alliance to Improve Outcomes in Acne: Large-scale worldwide observational study of adherence with acne therapy. International Journal of Dermatology **49**(4), 448–456 (2010). https://doi.org/https://doi.org/10.1111/j.1365-4632.2010.04416.x, `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-4632.2010.04416.x`
6. Grada, A., Muddasani, S., Fleischer, Jr, A.B., Feldman, S.R., Peck, G.M.: Trends in office visits for the five most common skin diseases in the united states. J. Clin. Aesthet. Dermatol. **15**(5), E82–E86 (May 2022)
7. Hayashi, N., Akamatsu, H., Kawashima, M., Acne Study Group: Establishment of grading criteria for acne severity. J. Dermatol. **35**(5), 255–260 (May 2008)

8. Huynh, Q.T., Nguyen, P.H., Le, H.X., Ngo, L.T., Trinh, N.T., Tran, M.T.T., Nguyen, H.T., Vu, N.T., Nguyen, A.T., Suda, K., Tsuji, K., Ishii, T., Ngo, T.X., Ngo, H.T.: Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. Diagnostics (Basel) **12**(8), 1879 (Aug 2022)
9. Inc., G.M.I.: Acne medication market trends analysis: Report 2023-2032. Tech. rep., Global Market Insights Inc. (2022), https://www.gminsights.com/industry-analysis/acne-medication-market
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), https://api.semanticscholar.org/CorpusID:6628106
12. Law, M.P.M., Chuh, A.A.T., Lee, A., Molinari, N.: Acne prevalence and beyond: acne disability and its predictive factors among chinese late adolescents in hong kong. Clin. Exp. Dermatol. **35**(1), 16–21 (Jan 2010)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
15. Shen, X., Zhang, J., Yan, C., Zhou, H.: An automatic diagnosis method of facial acne vulgaris based on convolutional neural network. Scientific Reports **8** (04 2018). https://doi.org/10.1038/s41598-018-24204-6
16. Strauss, J.S., Krowchuk, D.P., Leyden, J.J., Lucky, A.W., Shalita, A.R., Siegfried, E.C., Thiboutot, D.M., Van Voorhees, A.S., Beutner, K.A., Sieck, C.K., Bhushan, R., American Academy of Dermatology/American Academy of Dermatology Association: Guidelines of care for acne vulgaris management. J. Am. Acad. Dermatol. **56**(4), 651–663 (Apr 2007)
17. Tan, J.K.L., Bhate, K.: A global perspective on the epidemiology of acne. Br. J. Dermatol. **172 Suppl 1**, 3–12 (Jul 2015)
18. Tan, J.K.L., Fung, K., Bulger, L.: Reliability of dermatologists in acne lesion counts and global assessments. J. Cutan. Med. Surg. **10**(4), 160–165 (Jul 2006)
19. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
20. Wang, J., Luo, Y., Wang, Z., Hounye, A.H., Cao, C., Hou, M., Zhang, J.: A cell phone app for facial acne severity assessment. Appl. Intell. **53**(7), 7614–7633 (2023)
21. Wang, J., Wang, C., Wang, Z., Hounye, A.H., Li, Z., Kong, M., Hou, M., Zhang, J., Qi, M.: A novel automatic acne detection and severity quantification scheme using deep learning. Biomedical Signal Processing and Control

84, 104803 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.104803, https://www.sciencedirect.com/science/article/pii/S1746809423002367

22. Wen, H., Yu, W., Wu, Y., Zhao, J., Liu, X., Kuang, Z., Fan, R.: Acne detection and severity evaluation with interpretable convolutional neural network models. Technol. Health Care **30**(S1), 143–153 (2022)
23. Wu, X., Wen, N., Liang, J., Lai, Y.K., She, D., Cheng, M.M., Yang, J.: Joint acne image grading and counting via label distribution learning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10641–10650 (2019). https://doi.org/10.1109/ICCV.2019.01074
24. Yang, Y., Guo, L., Wu, Q., Zhang, M., Zeng, R., Ding, H., Zheng, H., Xie, J., Li, Y., Ge, Y., Li, M., Lin, T.: Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. Dermatology and Therapy **11** (05 2021). https://doi.org/10.1007/s13555-021-00541-9