Acne severity grading with deep learning

Abstract. Acne vulgaris is a skin condition which occurs frequently within the population. An important step in diagnosing the condition is grading the severity of the case. For this purpose dermatologists often use different grading scales and criteria. To aid this grading process, the usage of deep learning algorithms has been proposed by multiple authors. This paper explores the usage of deep learning algorithms on two separate datasets, one of which is reinforced with bounding boxes for acne.

Keywords: skin disorders · deep learning · acne assessment.

1 Introduction

Acne is a skin condition that can greatly impact quality of life for an individual in a negative way. Indeed, research shows that it can affect one's mental health greatly, especially among young people [20].

Nowadays the diagnosis of acne is done by the dermatologists who carefully observe skin of the patient to come up with conclusion. A lot of dermatologists are using so-called acne severity grading systems such as Global Acne Grading System (GAGS) [2] or Investigator Global Assessment of Acne (IGA) [3]. Severity grade is a helpful tool for dermatologists to select an appropriate treatment. It could also play a major role in designing clinical trials.

In this work we explore machine learning (ML) based approaches for automated acne grading. For this purpose, a dataset has been collected and labelled by a professional dermatologist with the grading criteria outlined. Along with the main dataset we use an additional one with annotated acne lesions, which allows applying semantic segmentation as well as object detection techniques.

The paper is organized as follows. In Section 2, we briefly summarise relevant research papers and projects. Section 3 describes our data and problem statement. Section 4 describes our experiments and obtained results. Finally, Section 5 concludes the paper and outline future prospects.

2 Related Work

ML methods have been recently successfully applied to different computer vision problems. Convolutional neural networks (CNNs) in particular have been showing great performances for such tasks as image classification [13], object detection [14] and semantic segmentation [9] among others.

This development has also brought advances to the medical domain. Work [7] explores different CNN architectures to build the model for detecting the presence of Lyme disease. In [4] authors apply Inception-v3 [18] architecture to detect diabetic retinopathy in retinal fundus photographs. As a result they achieve sensitivity and specificity scores of 97.5% and 93.4% correspondingly.

There are also several works that deal with acne specifically. Zhao et al. [23] claim to have developed a deep learning model which is capable of assessing acne severity from selfie images as accurately as dermatologists. They used transfer learning paradigm by extracting image features using a ResNet architecture pretrained model, then adding and training a fully connected layer to learn the target severity level from labeled images. Notably, authors also consider only four face areas of the original images (forehead, both cheeks, and chin) thus restricting the access of the model to the rest of the face. Work done by Zhang et. al [22] uses the ensemble approach solving classification and detection problems simultaneously. Classification block of the ensemble model uses ResNet architecture while the detection block is a you only look once (YOLO) [15] model. Finally, [19] train object detector models such as YOLOv4, faster region-based convolutional neural network (faster R-CNN) [16] with different backbones and single shot multibox detector (SSD) [12]. The work also uses predicted bounding boxes to evaluate severity by counting them.

3 Data and Problem Statement

Two datasets are used in this work. The first one, consists of 668 images of the so-called selfies (self-portraits of individuals usually taken with the help of a smartphone). The average resolution across the whole dataset is 931 by 674 ($H \times W$). The labels for this dataset are real numbers ranging from 0 to 1. They indicate the severity of acne, where the higher number indicates more severe case. More precisely, criteria used to obtain labels are shown in Table 1. To come to consensus regarding the criteria three dermatologists labeled image sets independently first, then analysis of their scoring differences was performed. To perform the analysis for each pair of dermatologists we built three figures. First one is constructed by sorting grades for both dermatologists according to one of them and plotting them in that order. Example is shown in Figure 1. The second Figure 3 demonstrates differences in form of the scatter plot, Pearson correlation score is 0.89. Finally, we compare distributions of their plots in Figure 2.

According to this criteria, the set of labels was obtained with the help of professional dermatologist. Distribution for the labels is shown in Figure 4. In order to fine-tune and test our models we split this dataset into training, test and validation sets with 9:1:1 ratio. To measure the quality of the models, we evaluate their performances on the validation set with the mean absolute error (MAE) and symmetric mean absolute percentage error (sMAPE) metrics. sMAPE is defined as follows:

$$sMAPE(p,t) = \frac{100}{n} \sum_{i=1}^{n} \frac{|p_i - t_i|}{(|p_i| + |t_i|)/2},$$

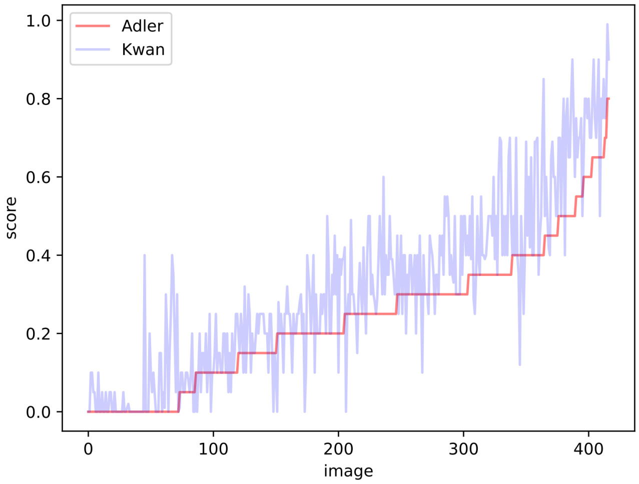

Fig. 1. Score differences: Dr. Adler vs. Dr. Kwan

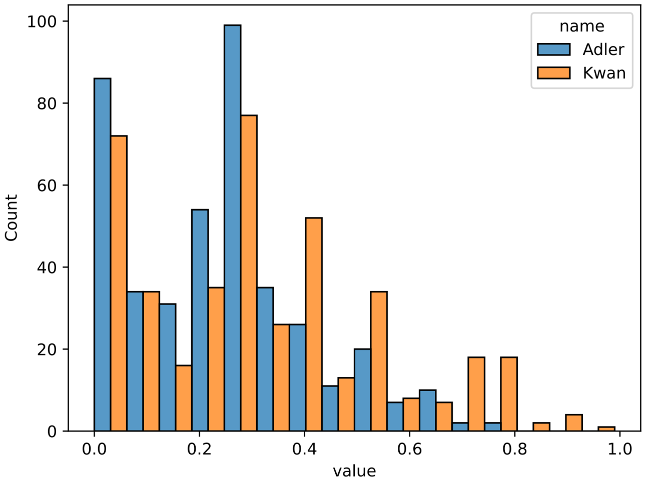

Fig. 2. Distribution differences: Dr. Adler vs. Dr. Kwan

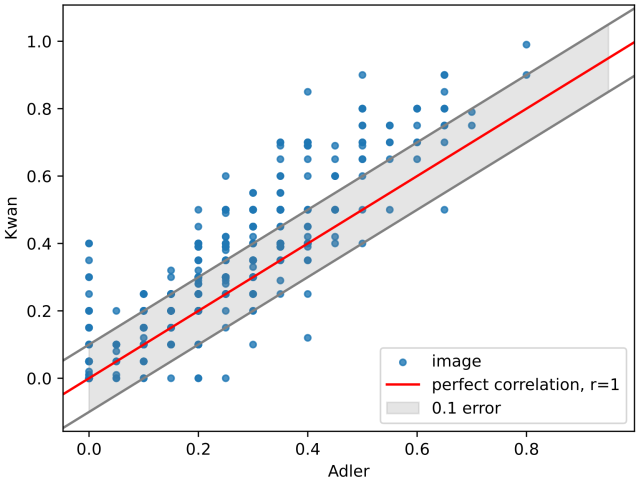

Fig. 3. Scatter plot: Dr. Adler vs. Dr. Kwan

Table 1. Grading criteria

Scores	Verdict	Criteria
0	No acne	- No acne Present
[0.01, 0.39]	Mild	- Lesion number is less in count - Distribution is localized
[0.04, 0.69]	Moderate	- Lesion number greater in count - Lesion size greater in size - May have a few nodules - Distribution is in more areas of the face than mild
[0.7, 1]	Severe	- Lesions, all over the face: pustules, nodules - Presence of 1 or more cystic acne lesions

where p is prediction vector, t truth vector and n is the size of both vectors. This quality metric is a symmetric version of MAPE and accounts for average relative error while MAE focuses on the average absolute error.

The second dataset, ACNE04, was proposed in [21]. Like the first one, it consists of face images. The difference is that it follows Hayashi's [5] requirements, so all images are taken at an approximately 70-degree angle from the front of patients. In total, there are 1457 images with the average resolution of 3027 by 2918 ($H \times W$). This dataset has both acne severity labels as well as bounding boxes of lesions annotated by professional dermatologists. There are 18 983 bounding boxes in total and their distribution is shown in Fig. 5. Severity grade is obtained from the lesion count. The criteria is provided in Table 2. For training purposes this dataset is already split with 8:2 ratio. The quality of the models trained on this dataset is evaluated via Intersection over Union (IoU) and Dice Coefficient for semantic segmentation and mAP@0.5 for object detection.

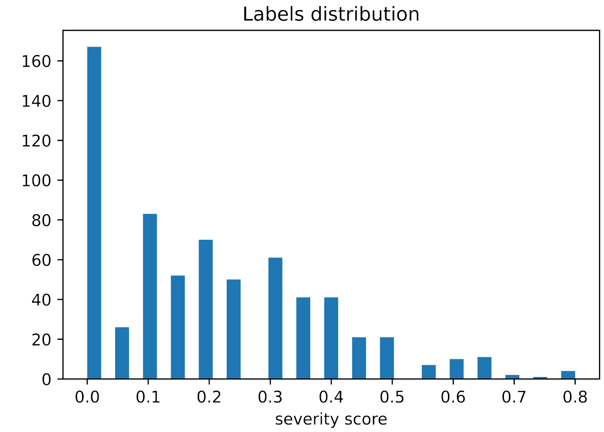

Fig. 4. Label distribution

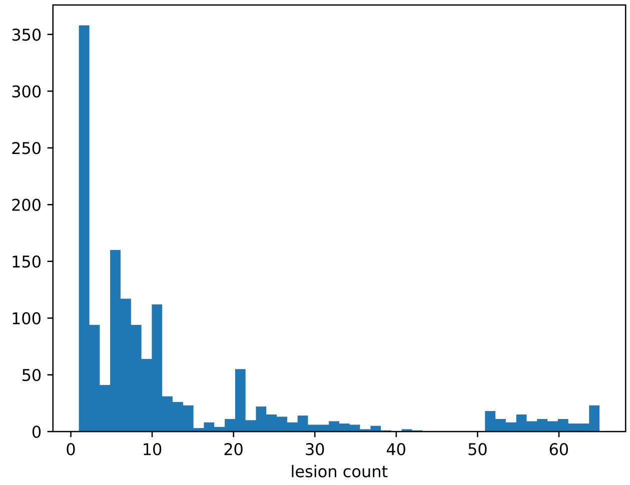

Fig. 5. Lesion count distribution

Table 2. Grading criteria 2

Class	Num. of lesions, l
Mild	$l \in [1, 5]$
Moderate	$l \in [6, 20]$
Severe	$l \in [21, 50]$
Very severe	$l > 50$

4 Experiments and Results

Initially, we use only the main dataset to solve the regression problem. For this, transfer learning approach was adopted. We use backbone CNN with pre-trained weights with a modified last layer to yield one real number from [0,1] interval. ResNet-18 [6] as well as MobileNet-v3 [8] were used as a backbone. For the loss function we choose MAE while the optimizer is Adam [10]. For the experiments an image processing pipeline was constructed. It includes resizing the images to 224 by 224 resolution, normalisation with mean and standard deviation of ImageNet [1] dataset. In order to increase the size of training data and robustness of the models, we also use augmentation techniques during the training procedures. It includes horizontal flips, Gaussian noise, small rotations and shifts. Pre-trained weights for both architectures were obtained on ImageNet. The resulting models however had underperforming metric scores with respect to acceptable quality level (see Table 4).

Next, we explore the approaches to localize acne before grading severity. We use ACNE04 data for this purpose. The starting approach was to use semantic segmentation to highlight acne lesions. To achieve this, we firstly transform our bounding boxes by making all pixels inside of them equal to 1 indicating target class zones. In this way, for each image in the additional dataset, we get a corresponding binary mask. We employ the U-Net model [17] next. For the backbone, we again test ResNet-18 and MobileNet-v3 pre-trained on ImageNet dataset. We preserve the same data processing pipeline as for the regression case with the only change being performed, i.e. resizing to higher resolution (480 by 480). For the loss function, the pixel-wise cross-entropy was used and the training phase was done with Adam optimizer. We observe that MAE scores on the validation are positive, while the SMAPE scores indicate that individually some predicted mask were far off. After the error analysis step, we note that the model performs better for severe cases, while the predictions for mild cases contains many false positives. Examples are provided in Fig. 6. The results are presented in table 4.

An alternative approach to localize acne was to use object detection techniques. We choose YOLO [15] model, namely, YOLOv8s implementation by ultralytics with the parameters pre-trained on the common objects in context (COCO) [11] dataset. The resizing was applied again to suit 640 by 640 COCO format. Resulting scores are presented in Table 3 and the prediction example is shown in Figure 7. We make use of the built model to solve the original regression problem on the main dataset. The approach is count based. Number of detected acne lesions by YOLO model were used as the only feature for linear regression. To count lesions we experimented with different confidence threshold for detected bounding boxes from 0 to 1 with step size 0.05 and choose the optimal one at 0.25. This resulted in an improvement (table 4) compared to the initial transfer learning approach. To improve further upon this we introduce the other factors such as the coverage and positioning. Coverage (C) is defined as a normalised total area taken by bounding boxes if treated as continuous rectangles:

$$C[i] = \frac{1}{H \times W} \sum_{j=1}^{M[i]} (b_i[j][x_{max}] - b_i[j][x_{min}]) \times (b_i[j][y_{max}] - b_i[j][y_{min}])$$

,where i is the current image, H and W are height and width of the image i, $M[i]$ is the amount of bounding boxes detected, $b_i[j]$ is the jth bounding box of image i and indices x_{max}, x_{min}, y_{max}, y_{min} indicate corresponding coordinates of the bounding boxes.

Positioning is defined as follows: we split the image into $n \times n$ grid and for each of the n^2 cells we count how many of detected bounding boxes fall into that cell. Bounding box b_i falls into the cell c_{rc} if their center of b_i is closer (Eucledian distance) to the center of c_{rc} compared to other cells. This way obtain n^2 new features related to relative positioning of bounding boxes. We the use of both Coverage and Positioning ($n = 2$) we obtain the improvement in performance (table 4).

Table 3. Results for ACNE04 dataset

Dataset	Backbone	Model	Metric	Score
ACNE04	ResNet18	U-net	IoU@0.2	0.16
ACNE04	ResNet18	U-net	Dice@0.2	0.28
ACNE04		YOLOv8s	mAP@50	0.33

Table 4. Results for original dataset

Dataset	Backbone	Model	MAE	SMAPE
Original	ResNet18	Backbone + FC linear layer	0.13	130
Original	MobileNetV3	Backbone + FC linear layer	0.09	84
Original		YOLOv8s + LR	0.08	64
Original		YOLOv8s + features + LR	0.078	63

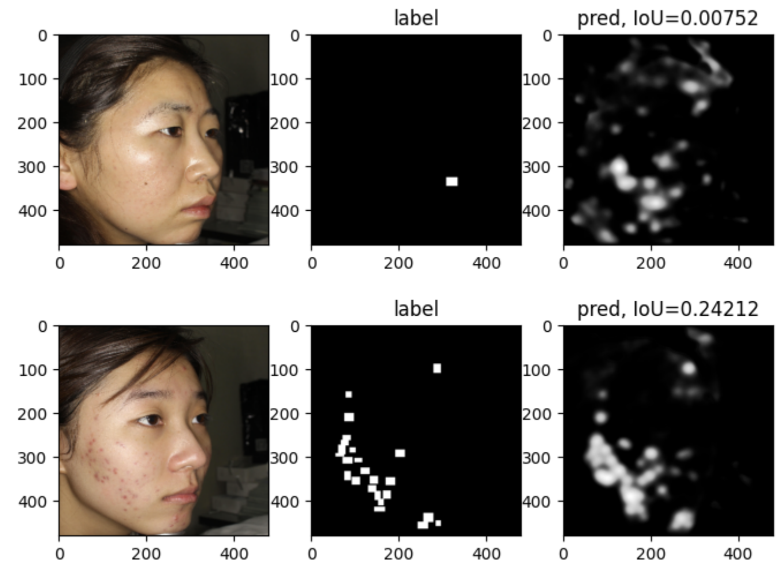

Fig. 6. U-net prediction examples

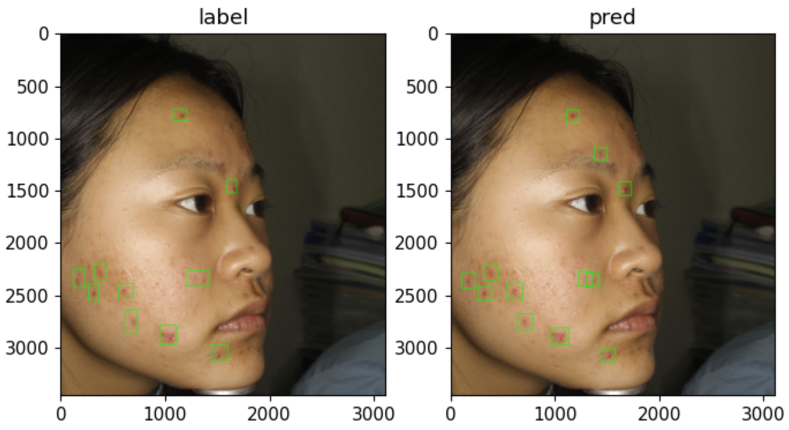

Fig. 7. YOLOv8 prediction examples

5 Conclusion and Future Work

As acne vulgaris continues to be a widespread skin condition worldwide the need for automatic sevirity grading grows.

In this study we explored the approaches to achieve automatic grading as precise as possible with respect to criteria from professional dermatologist. In the absense of the annotation for each acne lesion on the main dataset we use additional one ACNE04 in order to train acne detector first. With the use of YOLOv8s we achieved the mAP@0.5 score of 0.33 on the additional dataset. After obtaining the detector we use it to build the grader. Number of detected lesions and heuristic features such as coverage and positioning were proposed. With them we go from the original images data format to tabular dataset. We train linear regression on it and achieve MAE of 0.078 and SMAPE of 63%.

References

1. 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. Ieee (2009)
2. Doshi, A., Zaheer, A., Stiller, M.J.: A comparison of current acne grading systems and proposal of a novel system. Int. J. Dermatol. **36**(6), 416–418 (Jun 1997)
3. Food, Administration, D., et al.: Acne vulgaris: Establishing effectiveness of drugs intended for treatment. guidance for industry
4. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA **316**(22), 2402–2410 (12 2016). https://doi.org/10.1001/jama.2016.17216, https://doi.org/10.1001/jama.2016.17216
5. Hayashi, N., Akamatsu, H., Kawashima, M.: Establishment of grading criteria for acne severity. The Journal of dermatology **35**, 255–60 (06 2008). https://doi.org/10.1111/j.1346-8138.2008.00462.x
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
7. Hossain, S.I., de Goër de Herve, J., Hassan, M.S., Martineau, D., Petrosyan, E., Corbin, V., Beytout, J., Lebert, I., Durand, J., Carravieri, I., Brun-Jacob, A., Frey-Klett, P., Baux, E., Cazorla, C., Eldin, C., Hansmann, Y., Patrat-Delon, S., Prazuck, T., Raffetin, A., Tattevin, P., Vourc'h, G., Lesens, O., Nguifo, E.M.: Exploring convolutional neural networks with transfer learning for diagnosing lyme disease from skin lesion images. Comput. Methods Programs Biomed. **215**, 106624 (2022). https://doi.org/10.1016/j.cmpb.2022.106624, https://doi.org/10.1016/j.cmpb.2022.106624
8. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 1314–1324. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00140, https://doi.org/10.1109/ICCV.2019.00140

9. Hu, X., Jing, L., Sehar, U.: Joint pyramid attention network for real-time semantic segmentation of urban scenes. Appl. Intell. **52**(1), 580–594 (2022). https://doi.org/10.1007/s10489-021-02446-8, https://doi.org/10.1007/s10489-021-02446-8
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
11. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 21–37. Springer International Publishing, Cham (2016)
13. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 11966–11976. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01167, https://doi.org/10.1109/CVPR52688.2022.01167
14. Pham, V., Nguyen, D., Donan, C.: Road damage detection and classification with yolov7. In: Tsumoto, S., Ohsawa, Y., Chen, L., den Poel, D.V., Hu, X., Motomura, Y., Takagi, T., Wu, L., Xie, Y., Abe, A., Raghavan, V. (eds.) IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022. pp. 6416–6423. IEEE (2022). https://doi.org/10.1109/BigData55660.2022.10020856, https://doi.org/10.1109/BigData55660.2022.10020856
15. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv (2018)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. p. 91–99. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015), http://arxiv.org/abs/1505.04597
18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.308, https://doi.org/10.1109/CVPR.2016.308
19. 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)
20. Williams, H.C., Dellavalle, R.P., Garner, S.: Acne vulgaris. The Lancet **379**(9813), 361–372 (2012)
21. Wu, X., Ni, W., Jie, L., Lai, Y.K., Cheng, Dongyu, S., Ming-Ming, Yang, J.: Joint acne image grading and counting via label distribution learning. In: IEEE International Conference on Computer Vision (2019)
22. Zhang, H., Ma, T.: Acne detection by ensemble neural networks. Sensors **22**(18), 6828 (2022). https://doi.org/10.3390/s22186828, https://doi.org/10.3390/s22186828

23. Zhao, T., Zhang, H., Spoelstra, J.: A computer vision application for assessing facial acne severity from selfie images. CoRR **abs/1907.07901** (2019), http://arxiv.org/abs/1907.07901