scientific reports

Check for updates

OPEN

Evaluation of an acne lesion detection and severity grading model for Chinese population in online and offline healthcare scenarios

Na Gao[1,4], Jiaping Wang[2,4], Zheng Zhao[1,4]✉, Xiao Chu[2,4], Bin Lv[2], Gangwen Han[1], Yuan Ni[2]✉ & Guotong Xie[2,3]✉

Accurate acne severity grading is crucial for effective clinical treatment and timely follow-up management. Although some artificial intelligence methods have been developed to automate the process of acne severity grading, the diversity of acne image capture sources and the various application scenarios can affect their performance. Therefore, it's necessary to design special methods and evaluate them systematically before introducing them into clinical practice. To develop and evaluate a deep learning-based algorithm that could accurately accomplish acne lesion detection and severity grading simultaneously in different healthcare scenarios. We collected 2,157 facial images from two public and three self-built datasets for model development and evaluation. An algorithm called AcneDGNet was constructed with a feature extraction module, a lesion detection module and a severity grading module. Its performance was evaluated in both online and offline healthcare scenarios. Experimental results on the largest and most diverse evaluation datasets revealed that the overall performance for acne severity grading achieved accuracies of 89.5% in online scenarios and 89.8% in offline scenarios. For follow-up visits in online scenarios, the accuracy for detecting the changing trends reached 87.8%, with a total counting error of 1.91 ± 3.28 for all acne lesions. Additionally, the prospective evaluation demonstrated that AcneDGNet was not only much more accurate for acne grading than junior dermatologists but also comparable to the accuracy of senior dermatologists. These findings indicated that AcneDGNet can effectively assist dermatologists and patients in the diagnosis and management of acne, both in online and offline healthcare scenarios.

Keywords Acne severity grading, Lesion detection, Deep learning, Systematic evaluation

Acne is a common chronic inflammatory skin disease in the hair follicle sebaceous gland unit[1]. The prevalence of acne is estimated at over 80% among adolescents and about 9.4% for all ages, the eighth most common disease worldwide[2,3]. According to the types and/or the quantity of acne lesions, researchers have proposed many guidelines to distinguish the severity of acne and standardize the treatment regimens[4,5]. Incorrect severity grading or untimely treatment may lead to not only physical injuries such as facial scars, but also psychological distress such as depression and low self-esteem[3]. Thus, grading the severity of acne is very important, as it is essential for choosing a suitable treatment and monitoring curative effects[6].

Generally, acne severity grading is achieved through visual assessment when an acne patient seeks medical attention at the hospital. The diagnosis depends on the experience of the dermatologists, resulting in inter- and intra-observer variability[7]. Furthermore, due to insufficient medical resources, it is hard for people living in remote regions to get access to high-quality offline healthcare services[8]. Therefore, an image-based method for automated acne analysis has attracted increasing attention, as it can make acne diagnosis and follow-up management more objective and convenient in both offline and online healthcare scenarios. In the offline

[1]Department of Dermatology, Peking University International Hospital, Beijing, China. [2]Ping An Technology, Shanghai, China. [3]Ping An Healthcare and Technology Company Limited, Shanghai, China. [4] These authors contributed equally: Na Gao, Jiaping Wang, Zheng Zhao and Xiao Chu. ✉email: zhaozheng_2007@126.com; niyuan442@pingan.com.cn; xieguotong@pingan.com.cn

scenario, the acne images are captured by professional equipment such as a digital camera and VISIA skin care system[9,10], while in the online scenario they can also be taken by a smartphone[7,11]. The diversity of acne image capture sources has led researchers to develop robust image processing methods for acne analysis and evaluate their performance in multiple application scenarios.

Currently, many models have been proposed to identify skin diseases[12–17], Naeem et al. presented a deep learning-based method named DVFNet for the detection of skin cancer from dermoscopy images[12]. Ayesha et al. adpoted a multi-model fusion technique to achieve multi-classification of skin cancer[13]. Moreover, there are several methods specifically designed to identify acne. Most proposed methods are suitable for handling a single analysis task or a single capture source for acne images within limited application scenarios[9–11,18–23]. Some traditional image processing algorithms have been employed to detect and count acne lesions[18,19], but these solutions often lack robustness in practical scenarios, particularly when there is a significant difference between the training dataset and the validation dataset. In recent years, deep learning-based methods have largely improved the performance of acne image analysis[9–11,20–23]. Yang et al. accomplished facial acne severity grading classification based on Chinese Guidelines and reached an average F1 value of 0.8 on clinical images collected by digital cameras[9]. Lim et al. reported an automated calculation algorithm of the Investigator's Global Assessment (IGA) scale and showed a best classification accuracy of 67% on frontal facial view photographs[10]. Kim et al. developed an automated algorithm to detect acne lesions by type and proved it can improve the accuracy of lesion detection and counting[20]. Ravi et al. explored different techniques to improve the accuracy of acne vulgaris classification and found that the highest accuracy obtained on the original dataset was 89% by the EfficientNet model[21]. Trivedi et al. assessed the efficacy of various advanced object detection models, including Faster R-CNN, R-FCN, SDD, and YOLO, for the automated categorization of acne lesions in dermatological images[22]. These studies only concentrated on single task like acne severity grading classification or acne lesion detection[9,10,20], and only included acne images captured by professional equipment. Seite et al. suggested a facial acne analysis algorithm including acne severity grading as well as lesion identification, and achieved an accuracy of 68% on smartphone images according to the Global Evaluation Acne (GEA) scale[7]. Huynh et al. introduced an intelligent system on smartphone images and achieved a grading accuracy of 85% according to the IGA scale[11]. These studies combined acne lesion detection and acne severity grading together[7,11], but the acne image source only focused on smartphone images which were suitable for online healthcare scenarios. In addition, almost all of the current research is retrospective, with no prospective clinical studies being conducted[9–11,18–21]. It is necessary to develop an intelligent algorithm which can manage multiple acne analysis tasks and multiple image capture sources, and then evaluate its performance in different application scenarios.

In this paper, we aimed to develop and evaluate a deep learning-based algorithm for simultaneously achieving acne lesion detection and severity grading (AcneDGNet). The algorithm first employed a vision transformer[23] to generate stronger global feature representations for acne images from different capture sources. Then, it detected four types of acne lesions using convolutional neural networks[24,25] based on the guidelines of the Acne Grading System (AGS)[5]. Finally, acne severity grading was determined by fusing global and regional lesion-aware feature representations. The performance of our algorithm was evaluated in both online and offline healthcare scenarios using different datasets. These datasets included cross-sectional and longitudinal acne images taken by smartphones for online consultation scenarios, as well as retrospective and prospective cross-sectional acne images using digital cameras and VISIA skin care system for offline hospital diagnosis scenarios. Our contributions are summarized as follows: We developed a deep learning-based model (AcneDGNet) to achieve acne lesion detection and severity grading simultaneously; We evaluated the performance and reliability of AcneDGNet in both online and offline scenarios; We demonstrated that AcneDGNet was not only much more accurate than junior dermatologists but also comparable to the accuracy of senior dermatologists (Fig. 1).

Results

Study sample analysis

We collected two public datasets called ACNE04[16] and AcneSCU[26,27] and three self-built datasets called AcnePA1 (Acne data from Ping An Good Doctor), AcnePA2 and AcnePKUIH (Acne data from Peking University International Hospital). All images in the ACNE04, AcneSCU, and AcnePA1 were randomly divided for model development (training data) and evaluation (testing data), while the images in the AcnePA2 and AcnePKUIH were only used for model evaluation (testing data). Table 1 outlines the specific data partitioning process for model development and evaluation. Figure 2a further illustrates the distribution of severity grades across all datasets.

For model development, a total of 1,597 acne images were used and 35,557 lesion-bounding boxes were marked cumulatively. Figure 2b indicates the proportion of images captured by different devices: digital cameras accounted for 37%, smartphones accounted for 49%, and VISIA system accounted for the remaining 14%.

For model evaluation, we selected five datasets for two distinct application scenarios. The online scenario testing data consisted of 118 images from AcnePA1 and 148 images from AcnePA2, both captured exclusively by smartphone. The offline scenario testing data contained 147 images from the ACNE04 taken by digital cameras, 56 images from AcneSCU collected by the VISIA system, and 91 images from AcnePKUIH captured by cameras. Figure 2c visualizes the distribution of capture sources for testing data. We evaluated the grading performance of AcneDGNet by using recall and precision for each acne grade and accuracy for the whole.

The performance of AcneDGNet in online scenarios

In online scenarios, we used the testing data selected from AcnePA1 and AcnePA2 to evaluate the acne severity grading performance of AcneDGNet. Figure 3a shows some typical examples of the detected results obtained by AcneDGNet. The left four subgraphs represent the correctly identified cases corresponding to different acne severity grades, while the right subgraph depicts an incorrectly identified case. The primary reason for the

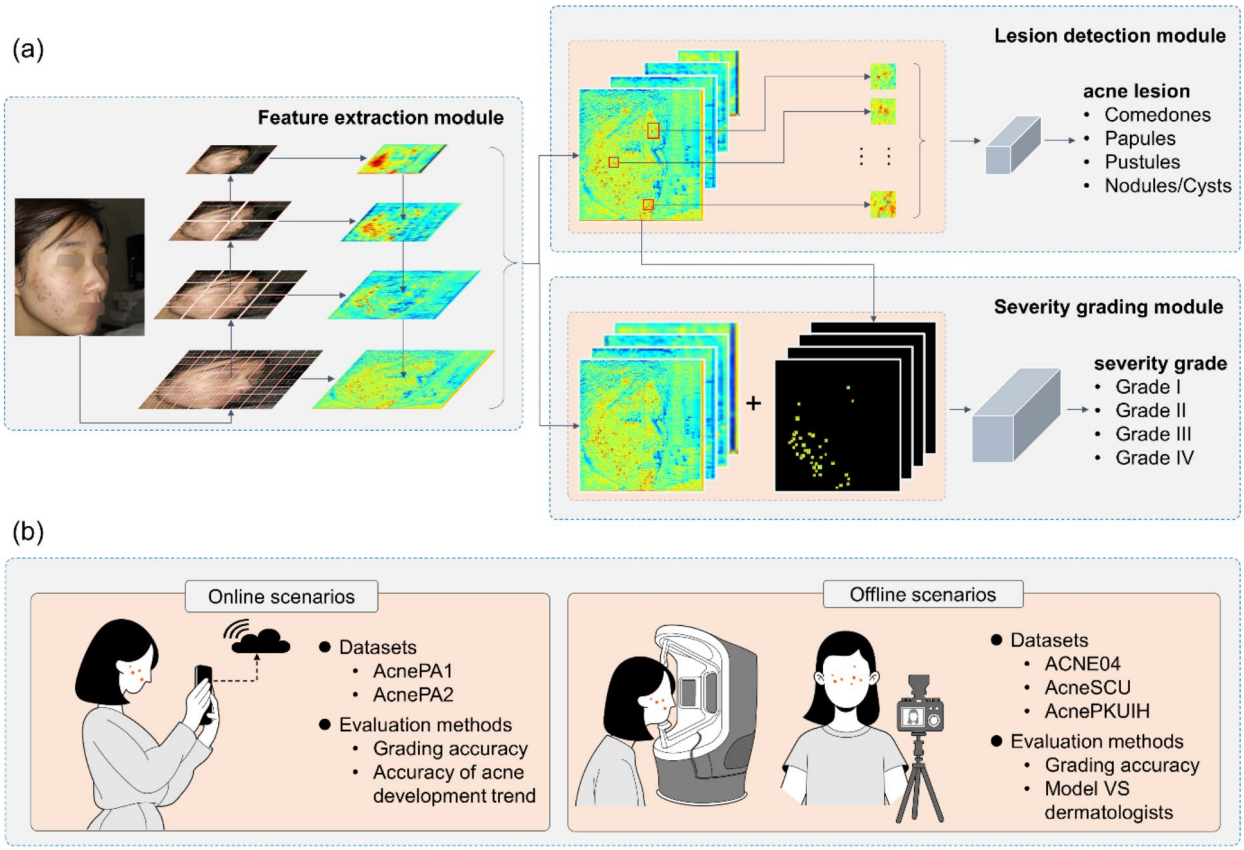

Fig. 1. The main framework of our proposed AcneDGNet and its evaluation. (**a**) is the framework of our proposed AcneDGNet. The input facial image is firstly passed through the feature extraction module, which includes Swin Transformer architecture and feature pyramid architecture. Then the multi-scaled feature maps output by the feature extraction module are input into the lesion detection module and the severity grade module, respectively. In the lesion detection module, the feature maps of acne candidate regions are obtained by region proposal network architecture to predict the location and category of each acne lesion in the image. In the severity grading module, the multi-scaled feature maps are resized and combine with the regional lesion-aware feature maps from lesion detection module to predict the severity grade of the acne image. (**b**) is the evaluation framework for the performance of AcneDGNet. We designed two application scenarios, including online scenario with smartphone and offline scenario with digital camera and VISIA system. Different datasets were selected for the corresponding evaluation purposes with different evaluation methods.

Purpose	Dataset	Capture source	Data distribution				
			Grade I	Grade II	Grade III	Grade IV	All
Model development	ACNE04[17]	Camera	20	377	92	97	586
	AcneSCU[22]	VISIA system	0	58	99	63	220
	AcnePA1	Mobile phone	56	486	91	158	791
Model evaluation for online scenario	AcnePA1	Mobile phone	7	81	7	23	118
	AcnePA2	Mobile phone	8	105	14	21	148
Model evaluation for offline scenario	ACNE04[17]	Camera	3	96	28	20	147
	AcneSCU[22]	VISIA system	0	15	30	11	56
	AcnePHUIH	Camera	4	63	11	13	91

Table 1. Data partitioning for all datasets.

inaccurate identification is that the AcneDGNet model misidentified water droplets as pustules. The results of detailed evaluation are summarized in Table 2, including recall and precision for each acne grade and the overall accuracy for all grades. The overall accuracy was 89.5% for AcnePA1&AcnePA2, while the individual accuracy was 85.6% for AcnePA1 and 92.6% for AcnePA2, respectively.

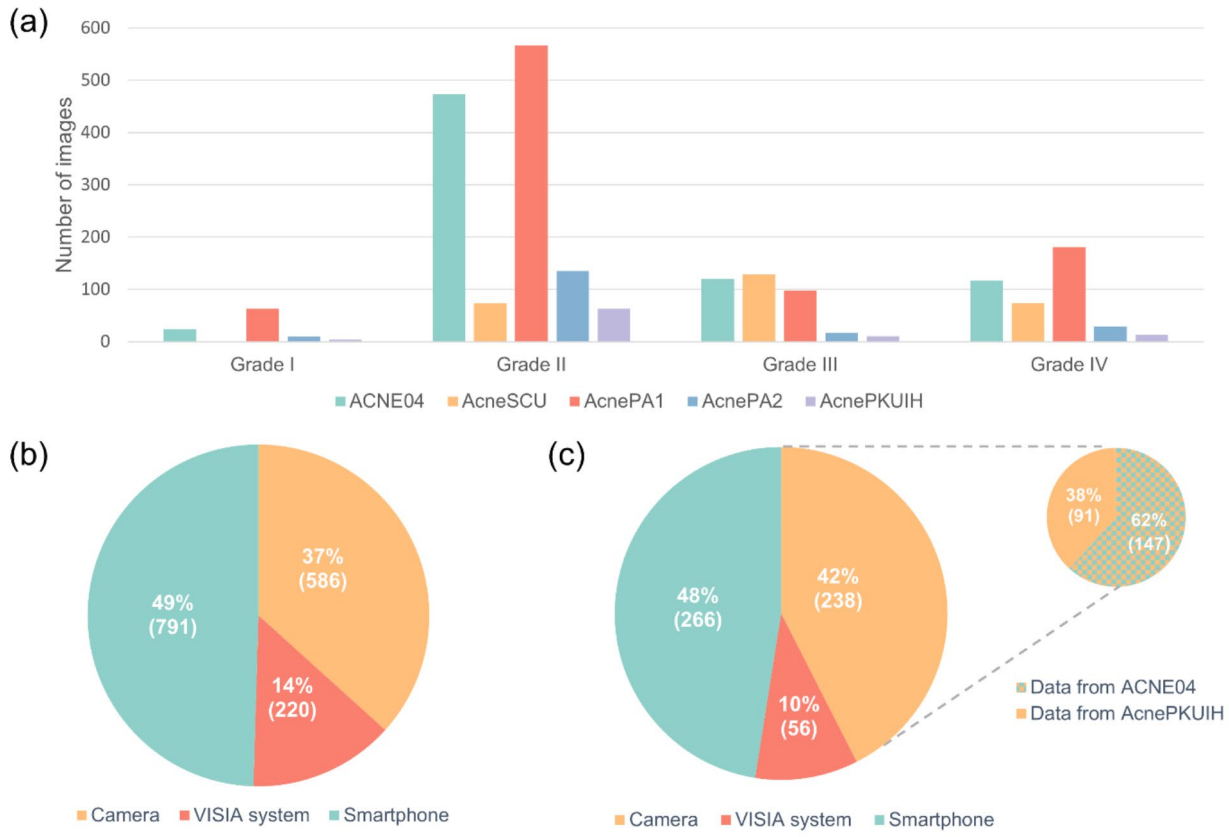

Fig. 2. The distribution of acne severity grading and the proportion of capture sources for all datasets. (**a**) is the distribution of image data volume at different acne severity grades for all datasets. (**b**) Is the proportion of capture sources for training data. And (**c**) is the proportion of capture sources for testing data, and the proportion of retrospective and prospective data which images were taken by camera.

Additionally, we evaluated the performance of AcneDGNet in detecting the changing trends of acne severity with regard to the grade and number of acne lesions. We applied AcneDGNet on AcnePA2 to predict acne grade on two images at different visit times for each patient and then determined whether the patient's acne grade was upgraded, unchanged or degraded. Figure 3b shows the detected results of AcneDGNet on two images for a patient during follow-up visits. Figure 3c is the confusion matrix for detecting the changing trends in the grade of acne severity. As shown in this figure, the overall accuracy was 87.8% since AcneDGNet correctly identified 65 cases of acne-changing trends among 74 follow-up visits. To evaluate the counting errors of the detected acne lesions, we calculated the mean and standard deviation of the absolute errors for the number of acne lesions between the AcneDGNet prediction and the ground truth. The total counting error was 1.91 ± 3.28 for all types of acne lesions. The individual counting errors were 0.70 ± 1.92 for comedones, 1.28 ± 2.01 for papules, 0.07 ± 0.29 for pustules, and 0.12 ± 0.38 for nodules/cysts, which can be seen in Fig. 3d.

The performance of AcneDGNet in offline scenarios

In offline scenarios, the grading performance of AcneDGNet was evaluated retrospectively based on ACNE04 and AcneSCU. Figure 4a,b present the detection results obtained by AcneDGNet on ACNE04 and AcneSCU, respectively. The detailed results of the retrospective evaluation are shown in Table 3. The overall accuracy across the entire retrospective dataset was 90.1%, with individual accuracies of 91.2% for ACNE04 and 87.5% for AcneSCU.

The grading performance of AcneDGNet was prospectively evaluated based on AcnePKUIH. Figure 4(c) shows the visualized results of AcneDGNet on AcnePKUIH. Moreover, we compared the model's performance to two junior dermatologists with over 5 years of experience (JD1 and JD2) and two senior dermatologists with over 10 years of experience (SD1 and SD2). The detailed results of the prospective evaluation are displayed in Table 4; Fig. 4(d). Specifically, the accuracy was 80.8% for two junior dermatologists, 90.7% for two senior dermatologists, and 89.0% for our AcneDGNet model, as shown in Table 4. The overall accuracy of AcneDGNet was 89.8% in offline scenarios by combining the retrospective and prospective datasets.

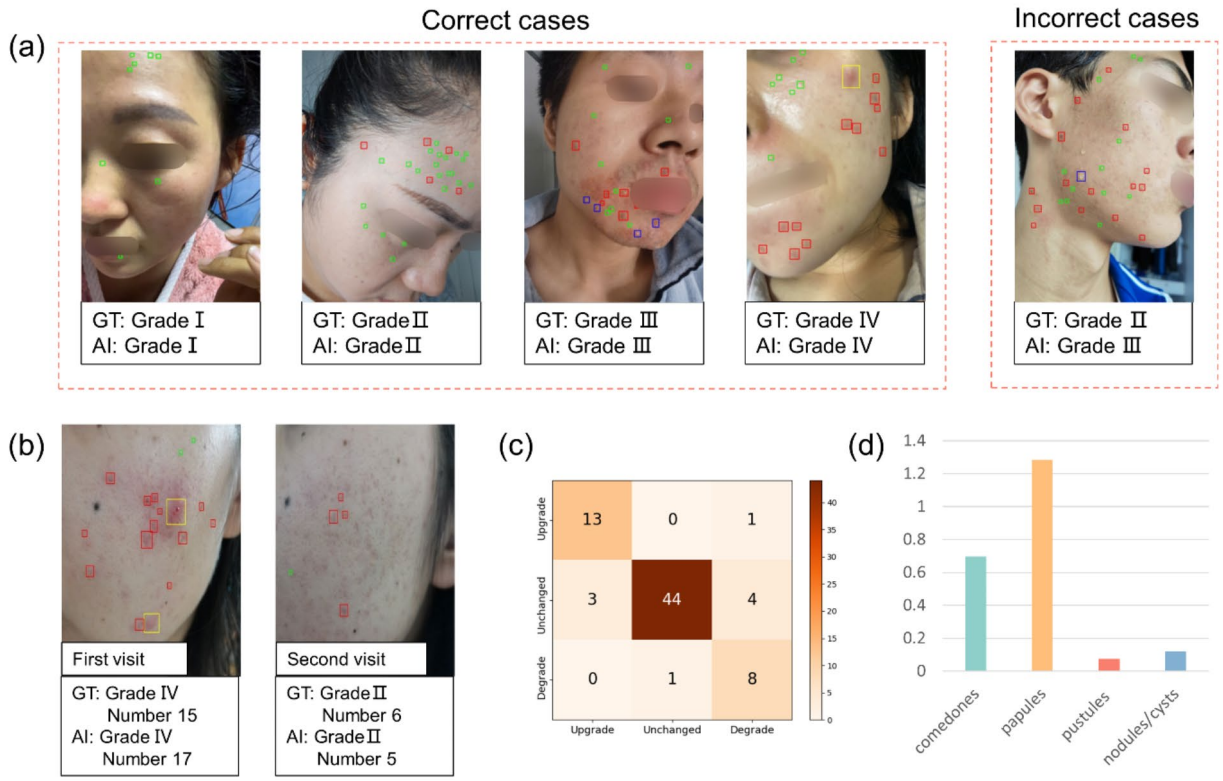

Fig. 3. The visualized results for the evaluation on online healthcare scenario. (**a**) shows some typical examples about identifying correct and incorrect cases by AcneDGNet on AcnePA1. The green, red, blue and yellow bounding boxes in the figure represent comedones, papules, pustules, and nodules/cysts, respectively. GT stands for Ground truth, which is the gold standard for acne grade. AI stands for Artificial Intelligence, which is the results predicted by our model.(**b**) is the visualized results of AcneDGNet on two follow-up images for a patient from AcnePA2. (**c**) is the confusion matrix for detecting the changing trends in the grade of acne severity. (d) is the counting error statistics of acne lesions for AcneDGNet regarding changing trends in the grade of acne on AcnePA2.

	AcnePA1			AcnePA2			AcnePA1&AcnePA2		
Acne grade	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)
Grade I	100.0	70.0	-	100.0	88.9	-	100.0	78.9	-
Grade II	87.7	93.4	-	91.4	100.0	-	89.8	97.1	-
Grade III	71.4	55.5	-	85.7	85.7	-	81.0	73.9	-
Grade IV	78.3	78.3	-	100.0	72.4	-	88.6	75.0	-
All grades	-	-	**85.6**	-	-	**92.6**	-	-	**89.5**

Table 2. The performance of AcneDGNet for acne severity grading in online scenario.

The performance comparison among AcneDGNet and previous studies

Table 5 shows the performance of our AcneDGNet compared to several previously similar studies on acne severity grading[9,11,26,28–30]. As shown in this table, our study had the most diverse data distribution across capture sources (smartphone, camera, and VISIA system) and the largest sample size of testing images from 486 patients. Our AcneDGNet achieved accuracies of 89.5% in the online scenario and 89.8% in the offline scenario, which outperformed the models from the most comparative studies[9,11,26,29,30] and was only outperformed by Acne-RegNet[28]. However, it should be noted that Acne-RegNet utilized multiple inputs, including several facial images and patient metadata information[28], which may account for its higher accuracy.

Discussion

We constructed AcneDGNet using a vision transformer and convolutional neural networks for acne lesion detection and severity grading (Fig. 1(a)). We evaluated its performance in both online and offline healthcare scenarios from multiple perspectives (Fig. 1(b)). The overall performance for acne severity grading achieved

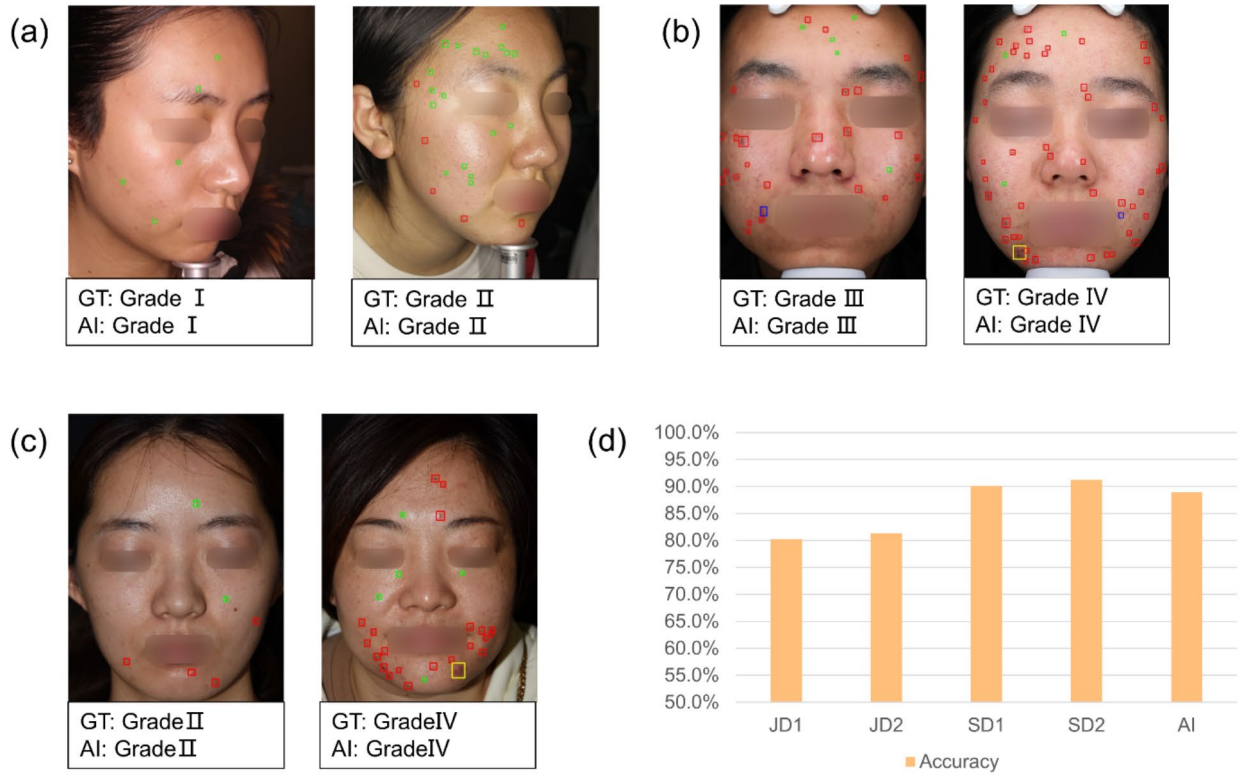

Fig. 4. The visualized results for the evaluation on offline healthcare scenario. (**a**) is an example of detected results of AcneDGNet on ACNE04, and (**b**) is another example of detected results of AcneDGNet on AcneSCU. (**c**) is an example from AcnePKUIH. And (**d**) shows the accuracy of four dermatologists (JD1, JD2, SD1, SD2) and AcneDGNet (AI) on AcnePHUIH, while JD represents junior dermatologist with more than 5 years' experience and SD represents senior dermatologist with more than 10 years' experience.

	ACNE04			AcneSCU			ACNE04&AcneSCU		
Acne grade	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)
Grade I	66.7	66.7	-	-	-	-	66.7	66.7	-
Grade II	97.9	93.1	-	100.0	78.9	-	98.2	90.8	-
Grade III	85.7	92.3	-	76.7	100.0	-	81.0	95.9	-
Grade IV	70.0	82.4	-	100.0	78.6	-	80.6	80.6	-
All grades	-	-	**91.2**	-	-	**87.5**	-	-	**90.1**

Table 3. The retrospective results of AcneDGNet for acne severity grading in offline scenario.

	Junior dermatologists (> 5 years' experience)			Senior dermatologists (> 10 years' experience)			Our AcneDGNet model		
Acne grade	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)	Recall (%)	Precision (%)	Accuracy (%)
Grade I	87.5	83.3	-	87.5	63.3	-	75.0	100.0	-
Grade II	78.6	94.5	-	92.1	94.3	-	95.2	89.6	-
Grade III	90.9	90.8	-	95.5	100.0	-	90.9	76.9	-
Grade IV	80.8	66.9	-	80.8	77.7	-	61.5	100.0	-
All grades	-	-	**80.8**	-	-	**90.7**	-	-	**89.0**

Table 4. The prospective results of human-machine performance comparison for acne severity grading in offline scenario.

Reference	Grading criteria	Capture source	Test patients	Model	Accuracy (%)
Yang et al.[10]	AGS scale[5]	Camera	40	Inception-v3	80.0
Wu et al.[17]	Hayashi criterion[7]	Camera	292	Label distribution learning	84.1
Lin et al.[16]	Hayashi criterion[7]	Camera	292	KIE/GLFNet	84.5
Liu et al.[25]	Hayashi criterion[7]	Camera	292	AcneGrader	85.8
Huynh et al.[12]	IGA scale[6]	Mobile phone	472	LightGBM	85.0
Wang et al.[15]	GAGS[26] and Hayashi criterion[7]	Mobile phone	147	Acne-RegNet	**94.56**
Ours	AGS scale[5]	Mobile phone/camera and VISIA system	192/294	AcneDGNet	**89.5/89.8**

Table 5. The performance comparison for acne severity grading among our AcneDGNet model and previous studies.

an accuracy of 89.5% in online scenarios and 89.8% in offline scenarios. For follow-up visits, the accuracy for detecting changing trends reached 87.8%, with a total counting error of 1.91 ± 3.28 for all types of acne lesions. The prospective evaluation in offline scenarios demonstrated that AcneDGNet was more accurate than junior dermatologists and was comparable to the accuracy of senior dermatologists (Table 4). Our experimental results on the largest and most diverse evaluation datasets showed that AcneDGNet had higher accuracy for acne severity grading than other state-of-the-art methods that used a single facial image as a model input (Table 5).

There are several medical guidelines for assessing acne severity. These include the Global Acne Grading System[31], the Hayashi criterion[6], the IGA scale suggested by US[32], the GEA scale encouraged in Europe, and the AGS scale recommended in China[4,5]. Among these guidelines, the AGS scale pays more attention to the type of acne lesion[5], while others incorporate both the type and quantity of acne lesions[4,6,32] or even consider the location of occurrence[31]. Recent researches have demonstrated that AGS was reliable for acne severity assessment and correlated well with other scales such as IGA and GEA[33,34]. All facial images in our study originate from the Chinese population, whether sourced from public datasets (ACNE04 [22] and AcneSCU[27]) or our self-built datasets (AcnePA1, AcnePA2, and AcnePKUIH). We therefore selected the AGS scale as the grading criteria since it might be more suitable for these datasets.

We made two special designs for AcneDGNet, one for feature extraction and the other for severity grading (Fig. 1). In the feature extraction module, we applied a vision transformer[23] to extract feature representations for acne lesion detection and acne severity grading. Compared to convolutional neural networks, the vision transformer had a stronger ability to extract the relationships between features across a wider range[23]. This module contributes to improving the generalization ability of the model among different image capture sources. In the severity grading module, we combined global feature maps with local lesion-aware feature maps. Most previous methods for acne grading only utilized either global features when using classification networks[9], or local features when using object detection networks[11,30]. A recent study proposed a framework called KIEGLFN to fuse global and local features[29]. However, its local features were extracted from local patches rather than acne lesions, leading to a lack of interpretability in the final results[29]. Our local feature maps were generated from the lesion detection module and were lesion-aware for different acne lesions. Due to these network designs, our AcneDGNet exhibited stable performance for acne severity grading in both online and offline scenarios (Tables 2 and 3, and 4), and achieved better performance than previous convolutional neural network-based models[9,11,26,29,30].

Traditionally, in offline scenarios, dermatologists made diagnoses based on the facial images collected by camera or VISIA[9,10] and other information during a face-to-face consultation. Previous studies have proposed some automatic methods for analyzing camera and VISIA images, achieving accuracy rates of 80–85.8% for acne severity grading[9,26,29,30]. In our study, the overall grading accuracy of AcneDGNet reached 89.8% in offline scenarios (Table 5). The retrospective result was 90.1% (Table 3), while the prospective result was 89.0% (Table 4). Our model achieved higher accuracy than junior dermatologists and was slightly lower than senior dermatologists. This indicates that our model could assist inexperienced dermatologists in assessing acne severity more accurately, especially in remote areas that lacked professional dermatologists.

Recently, online healthcare services have gained widespread attention and provided great convenience for patients and doctors, it's particularly suitable for acne diagnosis and follow-up management using a smartphone[11]. However, in online scenarios, image quality is highly uncontrollable, as patients may take facial images with varying shooting angles and complex backgrounds. One study reached an accuracy of 94.56% with Acne-RegNet[28]; However, its reliance on multiple facial images and systematic operational procedures made it inconvenient for patients in online scenarios. In our study, we collected real facial images through an online consultation platform (Ping An Good Doctor) and only selected one image as model input. Our AcneDGNet reached an accuracy of 89.5% for acne severity grading on the datasets from online consultations (Table 2). Furthermore, its performance for detecting the changing trends reached 87.8%, with a total counting error of 1.91 ± 3.28 for all acne lesions. This can provide effective assistance in follow-up visits (Fig. 1).

Certain limitations must be considered. Firstly, as displayed in Table 5, various studies[9,11,26,28–30] have utilized distinct grading criteria[5,6,31] and datasets with multiple capture sources[26,27]. We need to consider the potential impact of these differences while comparing their performances for acne severity grading. In the future, it would be ideal to compare the performance of various grading criteria on the same dataset. Secondly, our AcneDGNet can process facial images in a single modality. However, doctors often require multiple types of information for clinical diagnoses. The study by Wang, J. et al. has demonstrated that combining metadata information

with facial images can lead to improved accuracy[28]. Our next step is to develop a multi-modalities intelligence model to enhance the accuracy of acne diagnosis. Finally, image quality is one of the crucial factors influencing the stability of the model. Especially in online scenarios, the quality of the images uploaded by the patients is highly uncontrollable, and this may affect the performance of AcneDGNet. Further research about image pre-processing techniques or real-time user feedback during image capture is required while our AcneDGNet is applied to real-world conditions.

Conclusion

We developed a deep learning-based model called AcneDGNet to achieve acne lesion detection and severity grading based on the AGS scale and evaluated its efficiency in multiple application scenarios. Our AcneDGNet model is significant for grading acne severity, whether in online consultations or offline hospital diagnosis scenarios. The experimental results indicated that AcneDGNet could significantly assist doctors and patients in managing acne during consultations or routine follow-ups in clinical practice settings.

Methods

Ethics declarations

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Peking University International Hospital[2022-KY-0067]. Informed consent was obtained from all subjects and/or their legal guardian(s) to participate and publish in the study.

Image collection

We have collected two public datasets and three self-built datasets of facial acne images: (1) Two public datasets include ACNE04 and AcneSCU[26,27]. ACNE04 comprises 733 images that were collected by a digital camera[26], while AcneSCU comprises 276 images shot by the VISIA complexion analysis system[27]. (2) Three self-built datasets, AcnePA1 (Acne data from Ping An), AcnePA2, and AcnePKUIH (Acne data from Peking University International Hospital), were constructed using various methods. The images in AcnePA1 and AcnePA2 were taken by smartphones and uploaded by patients from the online consultation platform (Ping An Good Doctor). AcnePA1 comprises 909 cross-sectional images from 909 patients, while AcnePA2 comprises 148 longitudinal images from 74 patients. When building AcnePA2, we selected a pair of images with similar shooting angles from different visits for each patient. AcnePKUIH comprises 91 images from 91 patients, which were collected prospectively between December 2022 and April 2023 at Peking University International Hospital (PKUIH). The images were acquired using a Nikon d7200 DSLR camera. This study was approved by the ethics committee of PKUIH and adhered to the tenets of the Declaration of Helsinki. Written informed consent was obtained for all subjects before image collection.

In our initial quality control process, we eliminated the images according to the following criteria: (1) Images containing significant artifacts and noises; (2) Images taken so far away that acne lesions are too small to be seen; (3) Images that do not include any skin areas. We conducted anonymous processing and deidentification before transferring the images to the investigators. All images in our self-built datasets were processed by obscuring the eyes and mouth area to ensure privacy masking. Ultimately, we obtained 2,157 facial acne images for model development and evaluation.

Image labelling

According to the AGS scale proposed in the Chinese guidelines for managing acne[5], we classified the facial images into four clinical grades based on the type of acne. Specifically, Grade I means that only comedones occur, Grade II involves inflammatory red papules, Grade III indicates the presence of pustules and Grade IV includes nodules and/or cysts.

For the two public datasets ACNE04 and AcneSCU, as their original tasks were not the same as ours, we reviewed the labeled acne lesions and regraded the acne severity based on the AGS scale[5]. For the two self-built datasets AcnePA1 and AcnePA2, we labeled them from scratch. A team of four junior dermatologists with over 5 years of experience, two senior dermatologists with over 10 years of experience, and one dermatology expert with over 20 years of experience from PKUIH participated in the image annotation process. Before commencing mass labeling, these dermatologists underwent training to unify the annotation standard using a small set of acne images. Their Cohen's kappa score ultimately reached 0.71 on 100 randomly selected acne images. Then, the formal labeling process was conducted in three rounds. The entire dataset was randomly divided into four equal parts. In the first round, each quarter of the dataset was annotated independently by two junior dermatologists, which was equivalent to labeling two quarters of the dataset for each dermatologist. In this process, each dermatologist used LabelMe software to annotate acne lesions and grade acne severity based on the AGS scale[5]. Specifically, four types of acne were annotated in the form of bounding boxes with corresponding lesion types, including comedones, papules, pustules, and nodules/cysts. In the second round, two senior dermatologists reviewed together any inconsistent acne lesion annotations from the previous round. If there were still disagreements in the second round regarding acne grading, a dermatology expert would make the final decision in the third round.

For the prospective self-built dataset AcnePKUIH, the gold standard is determined by a dermatology expert in combination with clinical acne diagnosis records and images. Two junior dermatologists and two senior dermatologists independently performed acne grading labeling. These labels were utilized for human-machine performance comparison during the model evaluation phase.

Model development

We proposed a deep learning-based algorithm called AcneDGNet, which was specifically designed to perform acne lesion detection and severity grading simultaneously. The network architecture of AcneDGNet was constructed based on a vision transformer[23] and convolutional neural network[24]. The main functional modules of AcneDGNet include feature extraction, lesion detection, and severity grading modules, which are depicted in Fig. 1(a).

Within the feature extraction module, we applied Swin Transformer architecture[23] to obtain a more efficient hierarchical feature representation. Additionally, we utilized the feature pyramid architecture[25] to interact features across different scales. This integrated network structure exhibits stronger representation capabilities than a general convolutional neural network[35]. Subsequently, the extracted features were fed into the lesion detection module and severity grading module respectively. Within the lesion detection module, we employed a region proposal network architecture[24] to identify candidate acne regions from the feature maps. We then conducted more accurate localization and classification based on the features of these candidate regions. Ultimately, we obtained information regarding the location, category, and quantity of each acne lesion in the image. Within the severity grading module, we combined the global feature maps from the feature extraction module with the regional lesion-aware feature maps from the lesion detection module. This combination was utilized to establish a classification network for acne severity grading.

In the experiment, we initialized the AcneDGNet model using the pretrained weights on the ImageNet Dataset and then trained it on our collected acne data. To balance the model's accuracy and inference speed, we adopt the structural hyperparameters of Swin-S. Specifically, the channel number of the hidden layers in the first stage is 96, and the layer numbers are {2, 2, 18, 2}. To achieve stable model performance, we adopted the multi-scale training method mentioned in the Swin Transformer network paper[18]. Moreover, considering that acne lesions are generally small, we also appropriately increased the image size during training. Specifically, we randomly resized the input images such that the shorter side is between 680 and 1000 while the longer side is at most 1533. The cross-entropy loss and L1 loss were used as loss functions for lesion detection and severity grading. Because many acne lesions were too small to be detected, we especially replaced the bounding box regression loss function of region proposal network[24] in the lesion detection module with NWD-based regression loss function[36]. We used a learning rate of 0.0001 and the AdamW optimizer. The entire network was built using PyTorch 1.7.1, and training and testing were performed on an NVIDIA V100 GPU with CUDA version 10.2 cuDNN 7.6.5.

Model evaluation

To verify the effectiveness of our AcneDGNet model more comprehensively, we evaluated its performance in online and offline healthcare scenarios respectively, as shown in Fig. 1(b).

In the online consultation scenario, the acne images were taken and uploaded by smartphone. We selected two datasets, one was cross-sectional data from AcnePA1 for online diagnosis and another was longitudinal data from AcnePA2 for follow-up management. For the cross-sectional images, we primarily evaluated the grading performance of AcneDGNet with recall and precision for each acne grade.

$$Recall_i = \frac{TP_i}{TP_i + FN_i}, \ \ i = 1 \ldots \ N_{class} \tag{1}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, \ \ i = 1 \ldots \ N_{class} \tag{2}$$

where N_{class} represents the number of categories, TP/FP represents situations where positive samples are accurately/incorrectly identified by the AcneDGNet model, and FN represents situations where negative samples are incorrectly identified. In addition, we evaluated the overall accuracy based on the formulas as follows:

$$Accuracy = \frac{\sum_{i=1}^{N_{class}} TP_i}{N_{sample}} \tag{3}$$

where N_{sample} represents the total number of samples. For the longitudinal images, we also evaluated the performance of AcneDGNet regarding the changing trends in the number and grade of acne. The metrics were the same as above.

In the offline hospital diagnosis scenario, the capture sources for acne images were camera or VISIA system. Three datasets were selected to evaluate the grading performance of AcneDGNet, including two retrospective datasets (the data from ACNE04 [22] and AcneSCU[27]) and one prospective dataset (the data from AcnePKUIH). The metrics were the same as those mentioned above. Additionally, a human-machine performance comparison was conducted between AcneDGNet and four dermatologists using data from AcnePKUIH.

Data availability

The data that support the findings of this study are not publicly available due to the privacy of individuals that participated in the study, but are available on reasonable request. Zheng zhao should be contacted if someone wants to request the data from this study.

Received: 28 August 2024; Accepted: 25 December 2024
Published online: 07 January 2025

References

1. Tan, J. K. & Bhate, K. A global perspective on the epidemiology of acne. *Br. J. Dermatol.* **172**(Suppl 1), 3–12. https://doi.org/10.1111/bjd.13462 (2015).
2. Vos, T. et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the global burden of Disease Study 2010. *Lancet (London England)*. **380**, 2163–2196. https://doi.org/10.1016/s0140-6736(12)61729-2 (2012).
3. Sachdeva, M. et al. The prevalence, risk factors, and psychosocial impacts of acne vulgaris in medical students: A literature review. *Int. J. Dermatol.* **60**, 792–798. https://doi.org/10.1111/ijd.15280 (2021).
4. Dréno, B. et al. Development and evaluation of a global acne severity scale (GEA Scale) suitable for France and Europe. *J. Eur. Acad. Dermatology Venereology: JEADV*. **25**, 43–48. https://doi.org/10.1111/j.1468-3083.2010.03685.x (2011).
5. Chinese Guidelines for the Management of Acne Vulgaris. 2019 update #. *Int. J. Dermatology Venereol.* **2**, 129–138. https://doi.org/10.1097/jd9.0000000000000043 (2019).
6. Hayashi, N., Akamatsu, H. & Kawashima, M. Establishment of grading criteria for acne severity. *J. Dermatol.* **35**, 255–260. https://doi.org/10.1111/j.1346-8138.2008.00462.x (2008).
7. Seité, S., Khammari, A., Benzaquen, M., Moyal, D. & Dréno, B. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Exp. Dermatol.* **28**, 1252–1257. https://doi.org/10.1111/exd.14022 (2019).
8. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908. https://doi.org/10.1038/s41591-020-0842-3 (2020).
9. Yang, Y. et al. Construction and evaluation of a deep learning model for assessing Acne Vulgaris using clinical images. *Dermatology Therapy*. **11**, 1239–1248. https://doi.org/10.1007/s13555-021-00541-9 (2021).
10. V. Lim, Z. et al. Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin. Res. Technology: Official J. Int. Soc. Bioeng. Skin. (ISBS) [and] Int. Soc. Digit. Imaging Skin. (ISDIS) [and] Int. Soc. Skin. Imaging (ISSI)*. **26**, 187–192. https://doi.org/10.1111/srt.12794 (2020).
11. Huynh, Q. T. et al. Automatic acne object detection and Acne Severity Grading using smartphone images and Artificial Intelligence. *Diagnostics (Basel Switzerland)*. **12** https://doi.org/10.3390/diagnostics12081879 (2022).
12. Naeem, A., Anees, T. & DVFNet A deep feature fusion-based model for the multiclassification of skin cancer utilizing dermoscopy images. *PloS One*. **19**, e0297667. https://doi.org/10.1371/journal.pone.0297667 (2024).
13. Ayesha, H. et al. Multi-classification of skin Cancer using Multi-model Fusion technique. **5**, 195–219 (2023).
14. Naeem, A., Anees, T. J. J., o., C. & Informatics, B. A Multiclassification Framework for skin Cancer detection by the concatenation of Xception and ResNet101. 6, 205–227 (2024).
15. Naeem, A. et al. SNC_Net: Skin Cancer detection by integrating handcrafted and deep learning-based features using Dermoscopy images. **12**, 1030 (2024).
16. Naeem, A., Anees, T., Fiza, M., Naqvi, R. A. & Lee, S. W. SCDNet: a deep learning-based Framework for the multiclassification of skin Cancer using Dermoscopy images. *Sens. (Basel Switzerland)*. **22** https://doi.org/10.3390/s22155652 (2022).
17. Riaz, S., Naeem, A., Malik, H., Naqvi, R. A. & Loh, W. K. Federated and transfer learning methods for the classification of Melanoma and Nonmelanoma skin cancers: a prospective study. *Sens. (Basel Switzerland)*. **23** https://doi.org/10.3390/s23208457 (2023).
18. Maroni, G., Ermidoro, M., Previdi, F. & Bigini, G. in *2017 IEEE symposium series on computational intelligence (SSCI)*. 1–6 (IEEE).
19. Abas, F. S., Kaffenberger, B. & Bikowski, J. & Gurcan, M. N. in *Medical imaging 2016: computer-aided diagnosis*. 64–72 (SPIE).
20. Kim, D. H. et al. Automated facial acne lesion detecting and counting Algorithm for Acne Severity evaluation and its utility in assisting dermatologists. *Am. J. Clin. Dermatol.* **24**, 649–659. https://doi.org/10.1007/s40257-023-00777-5 (2023).
21. Ravi, K. M., Kiran, M. & Umadevi, V. in *1st International Conference on Communications and Computer Science (InCCCS)*. 1–6 (IEEE). (2024).
22. Trivedi, N. K. et al. in. *1st International Conference on Advanced Computing and Emerging Technologies (ACET)*. 1–5 (IEEE). (2024).
23. Liu, Z. et al. in *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
24. Ren, S., He, K., Girshick, R., Sun, J. & Faster, R-C-N-N. Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149. https://doi.org/10.1109/tpami.2016.2577031 (2017).
25. Lin, T. Y. et al. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
26. Wu, X. et al. in *Proceedings of the IEEE/CVF international conference on computer vision*. 10642–10651.
27. Zhang, J. et al. Learning High-quality Proposals for Acne Detection. (2022).
28. Wang, J. et al. A cell phone app for facial acne severity assessment. *Appl. Intell. (Dordrecht Netherlands)*. **53**, 7614–7633. https://doi.org/10.1007/s10489-022-03774-z (2023).
29. Lin, Y. et al. A unified acne grading framework on face images. *Comput. Methods Programs Biomed.* **221** https://doi.org/10.1016/j.cmpb.2022.106911 (2022).
30. Liu, S. et al. AcneGrader: an ensemble pruning of the deep learning base models to grade acne. *Skin. Res. Technology: Official J. Int. Soc. Bioeng. Skin. (ISBS) [and] Int. Soc. Digit. Imaging Skin. (ISDIS) [and] Int. Soc. Skin. Imaging (ISSI)*. **28**, 677–688. https://doi.org/10.1111/srt.13166 (2022).
31. Doshi, A., Zaheer, A. & Stiller, M. J. A comparison of current acne grading systems and proposal of a novel system. *Int. J. Dermatol.* **36**, 416–418. https://doi.org/10.1046/j.1365-4362.1997.00099.x (1997).
32. Food & Administration, D.
33. Yu, R., Fei, W., Ning, X. & Cui, Y. Inter-rater variability and consistency within four acne grading systems recommended in China, USA, and Europe. *J. Cosmet. Dermatol.* **21**, 6156–6162. https://doi.org/10.1111/jocd.15178 (2022).
34. Yu, R. et al. Comparisons of four Acne Grading systems recommended in China, Korea, and Japan. *Clin. Cosmet. Invest. Dermatology*. **16**, 203–210. https://doi.org/10.2147/ccid.S400226 (2023).
35. Zhang, D. et al. in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28*, Proceedings, Part XXVIII 16. 323–339 (Springer). (2020).
36. Wang, J., Xu, C., Yang, W. & Yu, L. J. a. p. a. A normalized Gaussian Wasserstein distance for tiny object detection. (2021).

Acknowledgements

The author(s) declare no competing interests.

Author contributions

Conceptualization: N.G., JP.W., Z.Z., X.C.; data curation: N.G., Z.Z.; formal analysis: N.G., JP.W., Z.Z., X.C., B.L., Y.N.; investigation: N.G., JP.W., Z.Z., X.C., B.L., Y.N.; methodology: JP.W.; project administration: GW.H., GT.X.; resources: GW.H., GT.X.; software: JP.W., X.C.; supervision: GW.H., Y.N., GT.X.; validation: N.G., B.L.; visualization: JP.W., B.L.; writing - original draft preparation: N.G., JP.W., Z.Z., X.C.; writing - review and editing: B.L., GW.H., Y.N., GT.X.; paper revision: JP.W.; Z.Z.; B.L.

Funding
Project supported by Peking University International Hospital Rearch Grant (No.YN2023QN11) and Science, Technology and Innovation Commission of Shenzhen Municipality (No.JSGG20210802153806021).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Z.Z., Y.N. or G.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025