

A Deep Learning-Based Automatic Object Detection and Classification Mechanism

Thi My Truong *, Seong Gon Choi *

* College of Information and Communication Engineering, Chungbuk National University, Chungdae-ro 1, Seowon-gu

mytruong@cbnu.ac.kr, choisg@cbnu.ac.kr

Abstract— Acne is a common skin disease that affects people of all ages and can result in hyperpigmentation, scarring, and negative psychological impacts. Therefore, detecting acne early and accurately diagnosing its severity is crucial for treatment and management. To address this issue, we proposed a system using a sequence of deep-learning models to automatically detect acne lesions and classify the severity of the global acne lesion existing in the face. Then, we propose a new grading algorithm to grade the final severity of acne lesions. Additionally, we develop a user-friendly application to help users who do not have expertise-knowledge can use it easily. It can serve as a support tool to assist dermatologists in making more precise diagnoses. Through experiments, our system has shown acceptable accuracy.

Keywords— deep learning, artificial intelligence, acne grading, object detection, classification.

I. INTRODUCTION

Acne is a chronic disorder usually associated with obstruction and inflammation of pilosebaceous units. It primarily affects areas of the skin with a higher density of oil glands, including the face, back, and chest. Although this disease is commonly found among adolescents, it is not restricted to this age group and can impact people of all ages. The severity of acne can differ wildly, from mild cases with just a few comedones to more severe forms marked by disfiguring inflammatory symptoms [1]. These severe cases may result in hyperpigmentation, scarring, and negative psychological impacts. Therefore, detecting acne early and accurately diagnosing its severity is crucial for treatment and management.

To diagnose acne, the patient must visit a dermatologist's clinic. The dermatologist then carefully examines the affected areas and uses additional information to provide a diagnosis [2]. This procedure is both time-intensive and reliant on the dermatologist's knowledge and clinical background. Furthermore, with a shortage of dermatologists in many regions globally, numerous patients are forced to travel far or endure extended wait times for diagnosis and treatment.

Owing to the rapid advancements in deep learning algorithms, artificial intelligence has become deeply integrated into the healthcare sector, particularly in areas that rely on morphological characteristics, such as dermatology [3]. Deep learning and convolutional neural networks can significantly enhance object detection, image classification, segmentation, and other related tasks. AI's extensive learning abilities enable it to identify subtle variations in lesion features, including number, size, texture, and finer details, frequently surpassing

human capabilities. This enables it to quantify lesion characteristics and provide assessments that aid in detecting and analyzing lesions, thereby improving the accuracy and efficiency of diagnoses.

In this paper, we propose the detection and assessment of acne severity using a deep learning architecture, and we also propose an application to diagnose facial acne based on a new severity grading algorithm. Key research contributions in this study include:

- We propose a system that uses a sequence of well-designed deep-learning models to detect acne and classify its severity.
- We proposed a new grading algorithm to assess the severity of acne based on the characteristics of the extracted acne lesions.
- Additionally, an app with a user-friendly interface has been designed and implemented to automatically diagnose acne and recommend treatments.

The structure of the manuscript is as follows: Section 1 presents an introduction of acne diagnosis and its progression through the application of deep learning techniques. Section 2 reviews the related work addressed in the study. Section 3 outlines the proposed methodology, while Section 4 details the experimental procedures. The results and discussion are provided in Section 5, and the conclusion is presented in Section 6.

II. RELATED WORK

In recent years, convolutional neural network-based approaches have gained significant traction in dermatology, particularly in the analysis of acne lesions, owing to the strong performance of deep learning in image processing tasks.

Zhao et al. [4] employed a transfer learning approach utilizing a pre-trained ResNet152 model to categorize acne severity into five distinct levels: clear (1), almost clear (2), mild (3), moderate (4), and severe (5). The performance of this model was benchmarked against the assessments of 11 dermatologists who evaluated the same acne cases.

Shen et al. [5] proposed a method that uses a 16-layer Visual Geometry Group (VGG16) to extract image features and classify facial acne vulgaris. It classified acne into various types, including non-inflammatory lesions and inflammatory lesions.

Yin Yang et al. [6] introduced a deep learning algorithm designed for grading facial acne severity based on the Chinese Guidelines. Their approach includes preprocessing facial images to exclude different face regions, followed by classifying acne lesions leveraging an Inception-V3 network.

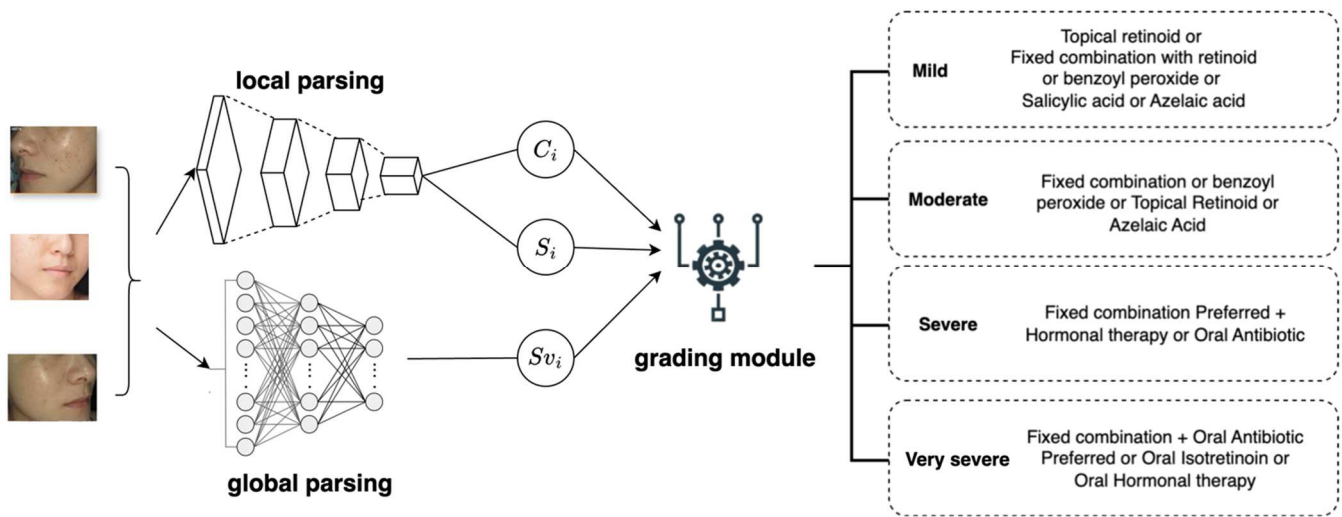


Figure 1. Overall System Architecture

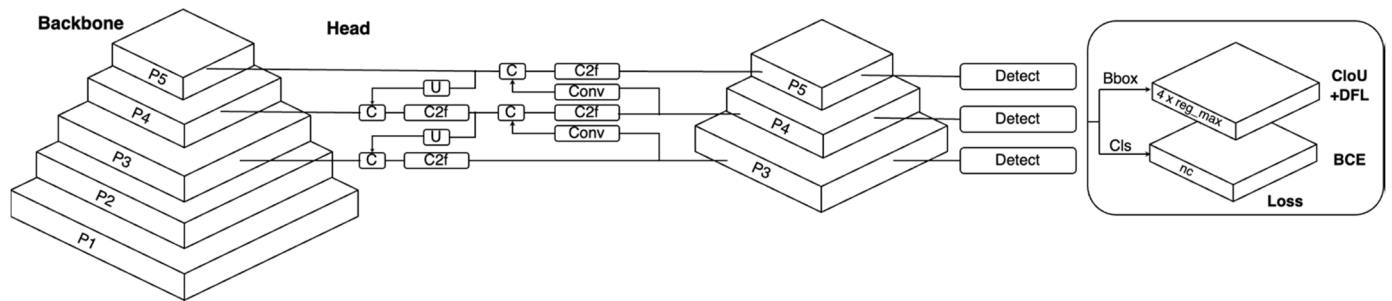


Figure 2. YOLOv8 model

The model's performance was evaluated on patients, achieving an average F1 score of 0.8.

Huynh et al. [7] designed an AI system to automatically detect acne and grade its severity using smartphone-captured facial images. It incorporates two models: a Faster R-CNN model for identifying four types of acne lesions and a LightGBM model that grades acne severity based on the Investigator's Global Assessment (IGA) scale using lesion counts provided by the Faster R-CNN. The Faster R-CNN model achieved a moderate mean Average Precision (mAP) of 0.54, while the LightGBM model demonstrated a high accuracy of 0.85 in severity grading.

III. PROPOSED METHOD

The pipeline of our proposed system is illustrated in Figure 1. The input consists of pictures taken from different angles. We pass these pictures through two branches. The first branch is local parsing to detect acne lesions, count the number, and calculate the size of each acne lesion. The second branch performs global parsing, classifying the acne lesion's severity. The characteristics are extracted:

- C is the total number of detected acnes in the input images.
- S_i is the size of the i -th acne lesion.

- S_V is the severity of the global acne lesion existing in the face.

These features are processed through a grading module that outputs the final severity results and recommends a treatment strategy. The details of each task are presented in the following subsection.

A. Detection model

Object detection is a computer vision method used to identify and label objects in images, video streams, or real-time footage. These models are trained on large datasets of annotated images, allowing them to accurately detect and classify objects in previously unseen data. We used the YOLOv8 model [8] for our acne detection task because of its proven efficiency, achieving state-of-the-art results with high mean average precision (mAP) scores and reduced inference time on the COCO dataset.

We trained the model using the ACNE04 dataset [9], which consists of 1,457 images. The training process encouraged the model to learn abstract feature representations from the data. The architecture of YOLOv8 is described in Figure 2. The backbone of the model, CSPDarknet53 [10], is a deep neural network designed to extract multi-scale features through a process of progressive down-sampling. These multi-resolution feature maps capture details about objects of varying scales

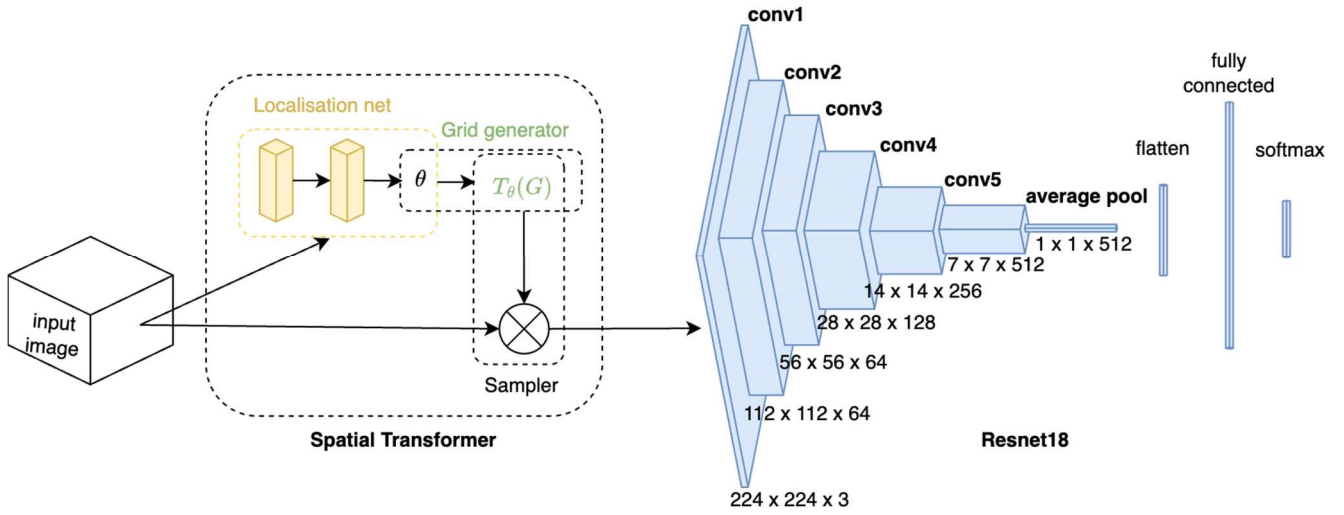


Figure 3. The Spatial Transform Resnet classifier.

within an image, facilitating the model's ability to learn object shapes and textures and contributing to its high performance in detection tasks.

CSPDarknet53's architecture consists of four sections, each featuring a convolutional layer followed by a c2f module. This module, newly introduced in CSPDarknet53, involves a split structure where part of the input passes through a bottleneck module comprising two 3x3 convolutions with residual connections. The output of this bottleneck module is split N times, where N reflects the size of the YOLOv8 model. These splits are concatenated and processed through a final convolution layer, which produces the activations for further layers in the network.

B. Classification model

The overall architecture of the classifier is illustrated in Figure 6. To boost the reliability of global parsing information, Spatial Transformer Network (STN) is employed in this work. STNs autonomously learn to focus on the most relevant facial regions affected by acne, effectively compensating for variations in facial positioning, rotation, and image quality. This capacity to zoom, rotate, and selectively attend to key areas allows the model to standardize inputs, thereby improving classification accuracy across a diverse range of patient images. Additionally, STNs facilitate the distinction of subtle differences in severity levels by concentrating on critical regions while also adapting to various scales of acne manifestation. Furthermore, STNs enhance the model's resilience to occlusions and mitigate dataset biases, resulting in improved generalization. By transforming input images into a more standardized format, STNs enable the primary classification network to process more consistent inputs, ultimately leading to more reliable and accurate severity assessments under diverse clinical imaging conditions.

1) Spatial Transformer Network: The non-ideal alignment of cropped bounding boxes within image frames can be attributed to spatial variations in the foreground, including

factors such as scale, rotation, and other geometric transformations. To address this, [11] introduced an approach that applies transformations directly to feature maps, enabling automatic operations such as cropping, scaling, and rotation. This work incorporates a STN to learn the necessary transformation parameters. The STN consists of three main components: a localization network, a grid generator, and a sampler.

The **Localization Network (LN)** serves as a straightforward CNN or MLP (Multi-Layer Perceptron), taking feature maps as input. In our implementation, the LN module was placed at the initial stage of the model, enabling direct processing of the input image by the LN module. The output of the LN is a matrix that encodes six parameters θ , which are used to compute the 2D affine transformation τ_θ . The spatial transformation is represented as $A_\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}$ indicating that the LN's output layer must consist of six nodes to represent these parameters.

The **Grid Generator** constructs a sampling grid based on the generated parameters θ . During this process, the Grid Generator iterates over the regular grid G of the target image, applying an inverse transformation T_θ to identify the corresponding locations within the input image. Specifically, each point in the input image is mapped to its corresponding position in the output image using a pointwise transformation governed by the previously learned transformation parameters, as described in equation (1).

$$\tau \left(\begin{bmatrix} x_{input} \\ y_{input} \end{bmatrix} \right) = A_\theta \times \begin{bmatrix} x_{output} \\ y_{output} \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \times \begin{bmatrix} x_{output} \\ y_{output} \\ 1 \end{bmatrix}, (1)$$

where the element 1 is appended to the homogeneous vector $\begin{bmatrix} x_{output} \\ y_{output} \end{bmatrix}$, it transforms the vector to represent a point in a

homogeneous coordinate system, thereby enabling its multiplication with the transformation matrix.

The **sampler** plays a crucial role in transforming the input image based on the learned transformation parameters. In the original study, the authors employed bilinear interpolation to achieve smoother output, demonstrating superior performance compared to other methods in practical evaluations. The mathematical formulation of bilinear interpolation is represented by the following equation:

$$f(x, y) = (1 - \alpha)(1 - \beta)f(x_1, y_1) + \alpha(1 - \beta)f(x_2, y_1) + (1 - \alpha)\beta f(x_1, y_2) + \alpha\beta f(x_2, y_2), \quad (2)$$

where:

- The values $f(x_1, y_1)$, $f(x_1, y_2)$, $f(x_2, y_1)$, $f(x_2, y_2)$ represent the four closest neighbour pixels.
- The parameter α denotes the proportional distance between the x-coordinate of the output pixel and its adjacent left pixel.
- The β quantifies the relative difference in the y-coordinate between the output pixel and its adjacent top pixel.

2) The Integrated Classifier: Our model utilizes Resnet18 [12] as the classification backbone. The network's convolutional layers predominantly use 3x3 filters, adhering to two key principles: (i) layers maintain the same number of filters if the output feature map size remains constant; and (ii) when the feature map size is reduced by half, the number of filters doubles, maintaining the time complexity per layer consistent. Downsampling is executed within convolutional layers featuring a stride of two. The architecture concludes with a global average pooling layer, followed by a 1000-way fully connected layer that incorporates a softmax function.

C. Grading module

In the assessment of acne severity, traditional methods often lack the nuance required to capture the complex nature of the condition. To address this limitation, we propose an advanced acne grading system that incorporates multiple factors extracted from multi-model operation. As aforementioned, the grading module in our automated acne assessment system utilizes three key factors derived from image analysis to quantify acne severity. These factors provide a comprehensive view of acne manifestation, combining both quantitative and qualitative aspects of skin condition assessment. The first factor, C , represents the total number of detected acne lesions in the input images. This count provides a fundamental measure of acne prevalence, offering insight into the extent of skin involvement. A higher value of C generally indicates a more severe case of acne, though this factor alone does not capture the full complexity of the condition.

Our second factor, S_i , denotes the size of the i -th acne lesion, where i ranges from 1 to C . This parameter allows us to consider not just the number of lesions, but also their individual magnitudes. Larger lesions often correlate with more severe or inflamed acne types, making this an essential component of our assessment. By incorporating S_i , we can differentiate between cases with numerous small lesions and those with fewer but larger, potentially more severe lesions.

The third factor, S_v , represents the severity score of the face generated by our classification system. This score encapsulates the overall appearance and condition of the skin, taking into account factors that may not be fully captured by lesion count and size alone. S_v provides a holistic assessment that can reflect aspects such as inflammation, redness, or texture changes associated with acne. To synthesize these factors into a cohesive grading system, we first calculate the average size of acne lesions. This average, S_{avg} , is computed as follows:

$$S_{avg} = \frac{1}{C} * \sum_{i=1}^C S_i, \quad (3)$$

By determining S_{avg} , we obtain a representative measure of typical lesion size for the patient. This allows us to account for variations in lesion size across different individuals and acne types. Next, we combine our three primary factors into a single, comprehensive Acne Severity Score (ASS). The formula for this score is:

$$ASS = (C * S_{avg}) + S_v, \quad (4)$$

The product of C and S_{avg} represents the total impact of acne lesions, considering both their number and size. By adding S_v to this product, we incorporate the overall severity assessment from our classification system, ensuring that our final score reflects both quantifiable lesion characteristics and broader skin condition factors. To translate our Acne Severity Score into clinically relevant categories, we employ a thresholding system. This system divides the continuous scale of ASS into distinct severity grades:

Grade = {
 “Mild” if $0 \leq ASS < T_1$,
 “Moderate” if $T_1 \leq ASS < T_2$,
 “Severe” if $T_2 \leq ASS < T_3$,
 “Very Severe” if $ASS \geq T_3$
 }

The threshold values T_1 , T_2 , and T_3 serve as boundaries between severity categories. These thresholds would be determined through careful analysis of clinical data and in consultation with experienced dermatologists. In our work, we empirically defined these thresholds as 25, 50, and 75, respectively.

IV. EXPERIMENTAL RESULTS

A. Environment setups

The models were implemented using the Pytorch framework on an Ubuntu 20.04 operation system with CUDA. The training process was conducted using the parameters outlined in Table I. To optimize the performance during training, the Adam optimizer was employed. The model was trained on a server equipped with an RTX 3090 GPU and an Intel Core i9-10980XE CPU.

TABLE 1. TRAINING PARAMETERS

Parameters	Value
Batch Size	32
Learning rate	4e-4
Momentum	0.9
Epoch	100

The UI was developed using the GTK toolkit version 4.15.7 for its graphical user interface. The code, written in C, was compiled with GCC version 11.4.0.

Following informed consent, each image was captured using a digital camera at a consistent 70-degree angle relative to the front of the patients. Professional dermatologists have annotated each image, marking acne lesions with bounding rectangles. This annotation process has resulted in a total of 18,983 bounding boxes across the dataset. The dataset also includes assessments of global acne severity, adhering to the Hayashi grading criterion, which provides overall severity classification from 0 to 3.

B. Results and Evaluation

1) Detection performance

It is evident that the outcomes of this phase significantly influence the subsequent stages of the processing pipeline. Using the recent state-of-the-art model YOLO v8 for training the dataset, this could help us to achieve a reliable result for our system. The results from training batch are shown as Figure 4. And the detection results shown in testing environments are depicted in Figure 5. The confusion matrix concludes 2 class and shown in the Figure 6. In limitation dataset and challenging small object of acne, it works with satisfied results in the scope of our study.



Figure 4. Training batch sample

2) Classification performance

TABLE 2. CLASSIFICATION PERFORMANCE

Models	Accuracy Train	Accuracy Test	FPS	PR(M)
ResNet18	0.91	0.80	309.8	11.6
STN-ResNet18	0.97	0.86	241.5	12.5

The comparison between the standard ResNet18 and the proposed STN-ResNet18 model demonstrates the significant benefits of incorporating a Spatial Transformer



Figure 5. Prediction test

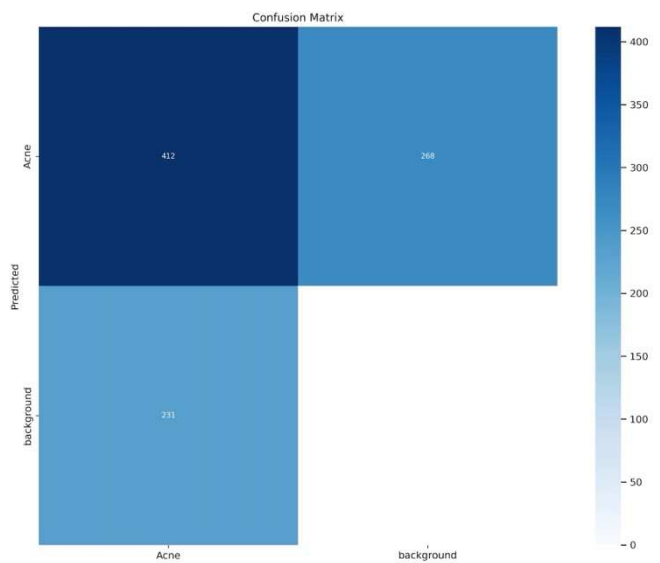


Figure 6. Confusion matrix

Network. STN-ResNet18 achieves superior accuracy in both training (0.97 vs 0.91) and testing (0.86 vs 0.80) scenarios, showcasing a remarkable 6 percentage point improvement in generalization capability. This substantial enhancement in accuracy is achieved with only a modest increase in model parameters (12.5M vs 11.6M), indicating an efficient use of additional complexity. While there is a trade-off in processing speed (241.5 vs 309.8 FPS), the 22% reduction in FPS is outweighed by the considerable gain in accuracy for many applications where precision is paramount. The STN-ResNet18 model's ability to dynamically focus on the most relevant parts of the input likely contributes to its superior performance, making it particularly well-suited for tasks requiring fine-grained visual analysis or dealing with varying object positions and scales. This innovative approach demonstrates that thoughtful architectural modifications, like incorporating an

STN, can yield significant improvements in deep learning model performance, potentially opening new avenues for enhanced accuracy in computer vision tasks.

3) System performance

The presented acne analyzer system demonstrates an innovative application of artificial intelligence in dermatological diagnostics shown in Figure 6. By offering a user-friendly interface for image capture and providing automated analysis of acne severity with corresponding treatment recommendations, this system shows potential to enhance the efficiency and accessibility of dermatological care. The classification of acne severity (e.g., "Mild") and tailored treatment suggestions indicate a sophisticated underlying algorithm. While the system's ability to quickly assess acne conditions and offer preliminary guidance is promising, it's important to view it as a supportive tool rather than a replacement for professional medical evaluation. Future research should focus on validating the system's accuracy across diverse skin types and conditions, assessing its impact on patient outcomes, and exploring its integration into clinical workflows.

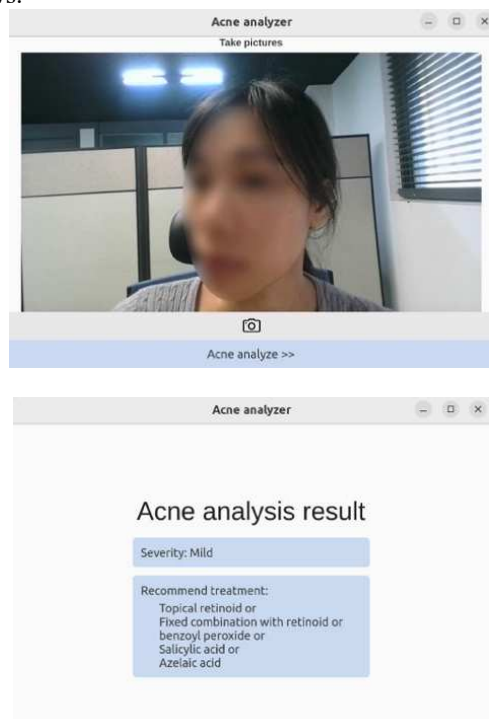


Figure 7. Application User Interface

V. CONCLUSION

This study emphasizes the potential of deep learning technologies to enhance the diagnosis and severity assessment of acne, representing a viable alternative to conventional diagnostic approaches. By developing an advanced deep-learning framework and introducing a novel severity grading algorithm, this work establishes a system that not only elevates the precision of acne detection but also broadens access to dermatological care via an intuitive application. Although the

findings are promising, they warrant additional validation in clinical environments and advocate for a broader diversity in the dataset to improve the system's robustness. This investigation contributes to the foundational work for future advancements in AI-powered healthcare, potentially improving patient outcomes and the efficiency of dermatological services significantly.

ACKNOWLEDGMENT

This work was partly supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2024-2020-0-01462, 50%) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945, 10%). This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00397979, Development of tera-level ultra-precision transmission network system technology, 40%)

*Corresponding author is S.G.Choi (choisg@cbnu.ac.kr)

REFERENCES

- [1] Sutaria, A. H., Masood, S., Saleh, H. M., & Schlessinger, J. (2023). Acne vulgaris. In *StatPearls [Internet]*. StatPearls Publishing.
- [2] Rashataprucksa, K., Chuangchaichatchavarn, C., Triukose, S., Nitinawarat, S., Pongprutthipan, M., & Piromsopa, K. (2020, August). Acne detection with deep neural networks. In *Proceedings of the 2020 2nd International Conference on Image Processing and Machine Vision* (pp. 53-56).
- [3] Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022). Artificial intelligence in dermatology image analysis: current developments and future trends. *Journal of clinical medicine*, 11(22), 6826.
- [4] Zhao, T., Zhang, H., & Spoelstra, J. (2019). A computer vision application for assessing facial acne severity from selfie images. *arXiv preprint arXiv:1907.07901*.
- [5] Shen, X., Zhang, J., Yan, C., & Zhou, H. (2018). An automatic diagnosis method of facial acne vulgaris based on convolutional neural network. *Scientific reports*, 8(1), 5839.
- [6] Yang, Y., Guo, L., Wu, Q., Zhang, M., Zeng, R., Ding, H., ... & Lin, T. (2021). Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. *Dermatology and therapy*, 11(4), 1239-1248.
- [7] Huynh, Q. T., Nguyen, P. H., Le, H. X., Ngo, L. T., Trinh, N. T., Tran, M. T. T., ... & Ngo, H. T. (2022). Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. *Diagnostics*, 12(8), 1879.
- [8] Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). Real-time flying object detection with YOLOv8. *arXiv preprint arXiv:2305.09972*.
- [9] Wu, X., Wen, N., Liang, J., Lai, Y. K., She, D., Cheng, M. M., & Yang, J. (2019). Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10642-10651).
- [10] Redmon, J. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- [11] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. *Advances in neural information processing systems*, 28.
- [12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).



Thi My Truong received B.S. degree in College of Information & Communication Engineering from Chungbuk National University in 2023. She is currently pursuing the Master degree in Radio Communication Engineering, Chungbuk National University. Her research interests include cybersecurity, Blockchain, computer vision.



Seong Gon Choi received B.S. degree in Electronics Engineering from Kyungpook National University in 1990, and M.S. and Ph.D. degree from KAIST in Korea in 1999 and 2004, respectively. He is currently a professor in College of Electrical & Computer Engineering, Chungbuk National University. His research interests include V2X, AI, smart grid, IoT, mobile communication, high-speed network architecture and protocol.